

AD-A086 521

NAVAL POSTGRADUATE SCHOOL MONTEREY CA
CLASSIFICATION TECHNIQUES FOR MULTIVARIATE DATA ANALYSIS. (U)
MAR 80 J K LEE

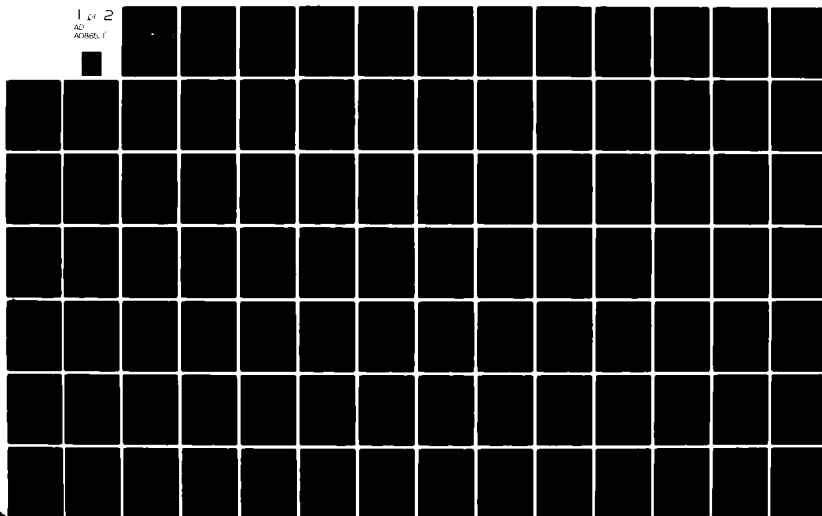
F/0 12/1

UNCLASSIFIED

NL

1 of 2

AD
A0865.1



LEVEL II

②

NAVAL POSTGRADUATE SCHOOL
Monterey, California

ADA 086521



SDTC
ELECTE
JUL 15 1980
D

THESIS

CLASSIFICATION TECHNIQUES FOR
MULTIVARIATE DATA ANALYSIS

by

Jin Ki Lee

March 1980

Thesis Advisor:

F. R. Richards

Approved for public release; distribution unlimited.

DDC FILE COPY.

80 7 14 108

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
	AD-A086521	
4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED	
6 Classification Techniques for Multivariate Data Analysis.	9 Master's Thesis March 1980	
7. AUTHOR(s)	6. PERFORMING ORG. REPORT NUMBER	
10 Jin Ki Lee		
8. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
Naval Postgraduate School Monterey, California 93940		
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE	
Naval Postgraduate School Monterey, California 93940	11 28 March 1980	
13. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	13. NUMBER OF PAGES	
Naval Postgraduate School Monterey, California 93940	120	
14. DISTRIBUTION STATEMENT (of this Report)	15. SECURITY CLASS. (of this report)	
Approved for public release; distribution unlimited.	Unclassified	
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. SUPPLEMENTARY NOTES		
251450 Jh		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
multivariate analysis, principal components analysis, variance-covariance matrix, correlation, lagrangian technique discriminant analysis, distance measure (metric), hierarchical clustering, nonhierarchical clustering, similarity matrix		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)		
<p>The multivariate analysis techniques of cluster analysis, principal components analysis, and discriminant analysis are examined in this thesis. The theory and applications of each of the techniques are discussed. Computer software available at the Naval Postgraduate School is discussed and sample jobs are included.</p> <p>A hierarchical cluster analysis algorithm, available in the IMSL software package, is applied to a set of data extracted from a</p>		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

group of subjects for the purpose of partitioning a collection of 26 attributes of a weapon system into six clusters of super-attributes.

A nonhierarchical clustering procedure, principal components analysis, and discriminant analysis were all applied to a collection of data on tanks considering of twenty-four observations of ten attributes of tanks. The cluster analysis shows that the tanks cluster somewhat naturally by nationality. The principal components analysis and the discriminant analysis show that tank weight is the single most important discriminator among nationality.

Accession For	
NTIS OVA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Available, or Special
A	

DD Form 1473
1 Jan 73
S/N 0102-014-6601

UNCLASSIFIED

2 SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Approved for public release, distribution unlimited

Classification Techniques for
Multivariate Data Analysis

by

Jin Ki Lee
Lieutenant Colonel, Korean Army
B.S., Korean Military Academy, 1966

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1980

Author

Lee Jin Ki

Approved by:

F. K. Richard

Thesis Advisor

Donald R. Barr

Co-Advisor or Second Reader

Michael J. Freeman

Chairman, Department of Operations Research

H. A. Schrag

Dean of Information and Policy Sciences

ABSTRACT

The multivariate analysis techniques of cluster analysis, principal components analysis, and discriminant analysis are examined in this thesis. The theory and applications of each of the techniques are discussed. Computer software available at the Naval Postgraduate School is discussed and sample jobs are included.

A hierarchical cluster analysis algorithm, available in the IMSL software package, is applied to a set of data extracted from a group of subjects for the purpose of partitioning a collection of 26 attributes of a weapon system into six clusters of superattributes.

A nonhierarchical clustering procedure, principal components analysis, and discriminant analysis were all applied to a collection of data on tanks considering of twenty-four observations of ten attributes of tanks. The cluster analysis shows that the tanks cluster somewhat naturally by nationality. The principal components analysis and the discriminant analysis show that tank weight is the single most important discriminator among nationality.

TABLE OF CONTENTS

	PAGE
I. DISCUSSION OF MULTIVARIATE DATA ANALYSIS . . .	8
II. PRINCIPAL COMPONENTS ANALYSIS	12
III. DISCRIMINANT ANALYSIS	19
A. Introduction	19
B. Theory	20
IV. CLUSTER ANALYSIS	32
A. Origin and Theory	32
B. Measure of Distance	34
C. Hierarchical Clustering	41
D. Nonhierarchical Clustering	47
V. ANALYSIS OF MULTIVARIATE UTILITY DATA . . .	52
VI. ANALYSIS OF ARMY TANK DATA	62
A. Data Structure	62
B. Nonhierarchical Cluster Analysis of Tank Data	64
C. Principal Components Analysis	69
D. Discriminant Analysis	73
VII. CONCLUSION	80
APPENDIX A: Program List for Hierarchical Clustering	81
APPENDIX B: Program List for MIKCA	88
APPENDIX C: Program List for Principal Component Analysis	115
APPENDIX D: Program List for Discriminant Analysis .	117
BIBLIOGRAPHY	118
INITIAL DISTRIBUTION LIST	120

LIST OF TABLES

	PAGE
I. ILLUSTRATIVE DATA MATRIX	9
II. DATA MATRIX	55
III. SIMILARITY MATRIX FOR SUPERATTRIBUTE DETERMINATION	58
IV. SUPERATTRIBUTES	61
V. THE FINAL CLUSTER SOLUTIONS	68

LIST OF FIGURES

	PAGE
1. EUCLIDEAN DISTANCE	38
2. TREE FOR HIERARCHICAL CLUSTERING	42
3. LOWER TRIANGULAR SIMILARITY MATRIX	43
4. TREE FOR 26 ATTRIBUTES	59
5. MIKCA FLOW CHART	65
6. SUMMARY TABLE OF PRINCIPAL COMPONENTS ANALYSIS ON TANK	70
7. GRAPHICAL PRESENTATION	71
8. FACTOR SCORE COEFFICIENTS	72
9. SUMMARY TABLE OF DISCRIMINANT ANALYSIS ON TANK .	75
10. CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS. .	76
11. SCATTER PLOT FOR ALL GROUPS	78

I. DISCUSSION OF MULTIVARIATE DATA ANALYSIS

A. INTRODUCTION

As a set of statistical techniques, multivariate data analysis is concerned with data collected on several dimensions of the same observations. Techniques can be used for many purpose in the behavioral, mathematical, and administrative sciences - ranging from rigidly controlled experiments to explain relationships assumed to be present in a large mass of data to attempts to cluster similar elements or to find functions of the variables that will best discriminate among preselected subpopulations of the observations.

The heart of any multivariate analysis consists of the data matrix. This matrix is a table that gives the results of a number of observations on a number of variables simultaneously (Table I).

Illustrative Data Matrix

Observations	Variables					
	1	2	3 j p	
1	x_{11}	x_{12}	x_{13}	x_{1j}	x_{1p}	
2	x_{21}	x_{22}	x_{23}	x_{2j}	x_{2p}	
.	
.	
.	
i	x_{i1}	x_{i2}	x_{i3}	x_{ij}	x_{ip}	
.	
.	
.	
n	x_{n1}	x_{n2}	x_{n3}	x_{nj}	x_{np}	

TABLE I.

The table consists of a set of observations (the n rows) and a set of measurements on those observations (the p columns). Cell entries represent the value x_{ij} of observation i on variable j . The values are characteristics of the observations and serve to define the observations in any specific study. The cell values may consist of nominal, ordinal, interval, or ratio-scaled measurements, or various combinations of these across columns.

In a general sense "multivariate" analysis would concern two main features:

1. The multivariate character lies in the multiplicity of the p variables, not in the size of the set n .
2. The variables are dependent among themselves so that we can not split off one or more from the others and consider it by itself. The variables must be considered together.

There are three characteristics often used as a basis for the classification of multivariate analysis:

1. whether one's principal focus is on the objects or on the variables of the data matrix;
2. whether the data matrix is partitioned into criterion and independent subsets, and the number of variables in each;
3. whether the cell values represent nominal, ordinal, or interval scale measurements.

This classification results in four major subdivisions of interest:

1. single criterion, multiple predictor association, including multiple regression, analysis of variance and covariance, and two-group discriminant analysis;

2. multiple criterion, multiple predictor association, including canonical correlation, multivariate analysis of variance and covariance, and multiple discriminant analysis;
3. analysis of variable interdependence, including factor analysis, multidimensional scaling, and other types of dimension-reducing methods;
4. analysis of interobject similarity, including cluster analysis and other types of grouping procedures.

The first two categories involve dependence structures where the data matrix is partitioned into criterion and independent subsets; in both cases interest is focused on the variables. The last two categories are concerned with interdependence - either focusing on variables or on observations. Within each of four categories, various techniques are differentiated in terms of the type of scale assumed.

In this research, we consider only the following techniques of multivariate analysis:

1. Principal components analysis
2. Discriminant analysis
3. Cluster analysis

II. PRINCIPAL COMPONENTS ANALYSIS

The basic idea of principal components analysis is to describe the dispersion of an array of n points in p -dimensional space by introducing a new set of orthogonal linear coordinates so that the sample variances of the given data points with respect to these derived coordinates are in decreasing order of magnitude. Thus the first principal component is such that the projection of given points onto it have maximum variance among all possible linear coordinates; the second principal component has maximum variance subject to being orthogonal to the first; and so on.

Suppose that the random variables X_1, X_2, \dots, X_p of interest have a certain multivariate distribution with finite mean vector u and variance-covariance matrix Σ .

From this population a sample of n independent observation vectors has been drawn. The observation can be written as the usual $n \times p$ data matrix.

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} X'_1 \\ \vdots \\ X'_n \end{bmatrix} \quad (1)$$

The estimate of Σ will be the usual sample variance-covariance matrix S defined as follows:

$$S = \frac{1}{n-1} A$$

$$A = \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})' \quad (2)$$

The information we shall need for our principal components analysis will be contained in S . However, it will be necessary to make a choice of measures of dependence: should we work with the variances and covariances of the observations, and carry out the analysis in original unit of the responses, or would a more accurate picture of the dependence pattern be obtained if each x_{ij} were transformed to a standardized score

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

and the correlation matrix R employed? The components obtained from S and R in general not the same, nor is it possible to pass from one solution to the other by a simple scaling of the coefficients.

If the responses are in widely different units (i.e., number of crew, weight in tons, speed in kilometer per hour, etc.) with large differences in the magnitudes, linear compounds of original quantities would have little

meaning and standardized variates and correlation matrix should be employed. Conversely, if the responses are reasonably commensurable, the covariance form has a greater statistical appeal, for the i -th principal component is that linear compound of the responses which explains the i -th largest portion of the total response variance, and maximization of such total variance of standard scores is rather artificial.

The first principal component of the complex of sample values of the responses X_1, X_2, \dots, X_p is the linear compound

$$Y_1 = a_{11}X_1 + \dots + a_{p1}X_p \quad (3)$$

whose coefficients a_{i1} are the elements of the eigenvector associated with the greatest eigenvalue λ_1 of the sample variance-covariance matrix of the responses. The a_{i1} are unique up to multiplication by a scale factor, and if they are scaled so that $a'_{i1}a_{i1} = 1$, the eigenvalue λ_1 is interpretable as the sample variance of Y_1 .

Numerical representation of the first principal component is to find the vector A_1 such that

$$\begin{aligned} Y_1 &= a_{11}X_1 + \dots + a_{p1}X_p \\ &= A'_1 X \end{aligned} \quad (4)$$

which maximizes sample variance

$$s_{Y_1}^2 = \sum_{i=1}^p \sum_{j=1}^p a_{i1} a_{j1} s_{ij} \quad (5)$$

$$= A_1' S A_1$$

for all coefficient vectors normalized so that $A_1' A_1 = 1$. To determine the coefficients, the normalization constraint is introduced by means of Lagrange multiplier and the resulting expression is differentiated with respect to A_1' :

$$\frac{\partial}{\partial A_1'} [s_{Y_1}^2 - \lambda_1 (1 - A_1' A_1)] = \frac{\partial}{\partial A_1'} [A_1' S A_1 + \lambda_1 (1 - A_1' A_1)] \quad (6)$$

$$= 2(S - \lambda_1 I) A_1$$

The coefficients must satisfy the p simultaneous linear equations.

$$(S - \lambda_1 I) A_1 = 0 \quad (7)$$

If the solution to these equation is to be other than the null vector, the value of λ_1 must be chosen so that

$$|S - \lambda_1 I| = 0 \quad (8)$$

λ_1 is thus an eigenvalue of the variance-covariance matrix, and A_1 is its associated eigenvector. To determine which of the p eigenvalues should be used, premultiply the

the system of equation (7) by A_1' . Since $A_1' A_1 = 1$, it follows that

$$\lambda_1 = A_1' S A_1 = s_{Y_1}^2$$

But the coefficient vector was chosen to maximize this variance, and therefore, λ_1 must be the greatest eigenvalue of S .

The second principal component is that linear compound

$$Y_2 = a_{12}X_1 + \dots + a_{p2} X_p \quad (9)$$

whose coefficients have been chosen, subject to the constraints

$$\begin{aligned} A_2' A_2 &= 1 \\ A_1' A_2 &= 0 \end{aligned} \quad (10)$$

so that the variance of Y_2 , $A_2' S A_2$, is a maximum. The first constraint is merely a scaling to assure the uniqueness of the coefficients, while the second requires that A_1 and A_2 be orthogonal.

The coefficients of the second component can also be found by the Lagrangian technique with two multipliers λ_2 and μ . Differentiating this with respect to A_2 gives:

$$\begin{aligned} & \frac{\partial}{\partial A_2} [A_2' S A_2 + \lambda_2 (1 - A_2' A_2) + \mu A_1' A_2] \\ & = 2(S - \lambda_2 I)A_2 + \mu A_1 \end{aligned} \quad (11)$$

If the right-hand side is set equal to 0 and premultiplied by A_1' , it follows from the normalization and orthogonality conditions that

$$2 A_2' S A_2 + \mu = 0 \quad (12)$$

Similar premultiplication of the equation (7) by A_2' implies that

$$A_2' S A_2 = 0 \quad (13)$$

and hence $\mu = 0$. The second vector must satisfy

$$(S - \lambda_2 I)A_2 = 0 \quad (14)$$

And it follows that the coefficients of the second component are thus the elements of the eigenvector corresponding to the second greatest eigenvalue. The remaining principal components are found in their turn in the same manner from the other eigenvectors.

Thus the j -th principal component of the sample of p -variate observations is the linear compound

$$Y_j = a_{1j}X_1 + \dots + a_{pj}X_p \quad (15)$$

whose coefficients are the elements of the eigenvector of the sample variance-covariance matrix S corresponding to the j -th largest eigenvalue λ_j . If $\lambda_i \neq \lambda_j$, the coefficients of the i -th and j -th components are necessarily orthogonal; if $\lambda_i = \lambda_j$, the elements can be chosen to be orthogonal, although an infinity of such orthogonal vectors exists. The sample variance of the j -th components is λ_j , and the total system variance is thus

$$\lambda_1 + \lambda_2 + \dots + \lambda_p = \text{tr } S \quad (16)$$

The importance of the j -th component in a more parsimonious description of the system is measured by

$$\frac{\lambda_j}{\text{tr } S} \quad (17)$$

which gives the fraction of the total variance contributed to the j -th component.

III. DISCRIMINANT ANALYSIS

A. INTRODUCTION

The basic idea of discriminant analysis consists of assigning an individual from a group of individuals to one of several known or unknown distinct populations, on the basis of observations on several characters of the individual or group and a sample of observations on these characters from the populations if these are unknown.

Fisher (1936) was the first to suggest a linear function of variables representing different characters, hereafter called the linear discriminant function (discriminator) for classifying an individual into one of two populations. Later research extended the analysis to classification into one of k populations.

For the univariate case Fisher suggested a rule which classifies an observation x into the i -th univariate population if

$$X - \bar{X}_i = \min (X - \bar{X}_1, X - \bar{X}_2), \quad i = 1, 2 \quad (18)$$

where \bar{X}_i is the sample mean based on a sample of size N_i from i -th population. For two p -variate populations π_1 and π_2 (with the same covariance matrix) Fisher replaced the vector random variable by an optimum linear combination of its components obtained by maximizing the

ratio of the difference of the expected values of a linear combination under π_1 and π_2 to its standard deviation. He then used his univariate discrimination method with this optimum linear combination of components as the random variable.

Rao (1948) considered the problem of classifying people into one of these populations castes of India. He assumed that each of the three populations could be characterized by four variables - structure (x_1), sitting height (x_2), nasal depth (x_3), and nasal height (x_4) - of each member of the population. On the basis of sample observations on these characters from the three populations the problem is to classify an individual with observation $X = (x_1, x_2, x_3, x_4)^T$ into one of three populations. He used a linear discriminator to obtain the solution.

B. THEORY

In general, the underlying assumptions of discriminant analysis are:

1. the groups being investigated are discrete and identifiable;
2. each observation in each group can be described by a set of measurements on p characteristics or variables;
3. these p variables are assumed to have a multivariate normal distribution in each population.

The purposes of discriminant analysis are:

1. to test for mean group differences and to describe the overlaps among groups;
2. to construct classification schemes based upon the set of p variables in order to assign previously unclassified observations to the appropriate groups.

Hence, the problem of studying the direction of group differences is, equivalently, a problem of finding a linear combination of the original independent variables that shows large differences in group means. In short, discriminant analysis is a method for determining such linear combinations.

The first step toward determining a linear combination of a set of variables such that several group means on this linear combination will differ widely among themselves, is to decide on a criterion for measuring such group-mean differences. Once a linear combination has been constructed, that means there is just a single transformed variable. Hence, the F -ratio for testing the significance of the over all difference among several group means on a single variable suggests an appropriate criterion.

$$F = \frac{v' B v}{v' W v} = \lambda \quad (19)$$

where $v' = (v_1, v_2, \dots, v_p)$, a set of weight which maximizes λ .

$$B = \sum_{i=1}^G N_i (\bar{x}_{i.} - \bar{x}_{..}) (\bar{x}_{i.} - \bar{x}_{..})'$$

$$W = \sum_{i=1}^G \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.}) (x_{ij} - \bar{x}_{i.})'$$

x_{ij} is the j th observation vector in the i -th group.

$\bar{x}_{..}$ is the grand mean vector of the data.

G is the number of groups.

n_i is the number of observations in the i th group.

Prime notation indicates transpose.

This ratio λ , called the discriminant criterion, was originally proposed by Fisher in connection with his two-group discriminant function. Once a criterion for group differentiation has been determined, a set of weights, (v_1, v_2, \dots, v_p) , which maximizes this criterion, should be determined. This is accomplished by taking the partial derivative of λ with respect to each component v_i of v and setting the result equal to zero.

$$\begin{aligned} \frac{\partial \lambda}{\partial v} &= \frac{2[(Bv)(v'Wv) - (v'Wv)(Wv)]}{(v'Wv)^2} \\ &= \frac{2(Bv - Wv)}{v'Wv} = 0 \end{aligned} \tag{20}$$

which is equivalent to

$$\begin{aligned}(B - \lambda W)v &= 0 \\ (W^{-1}B - \lambda I)v &= 0\end{aligned}\tag{21}$$

This equation is of the form

$$(A - \lambda I)v = 0\tag{22}$$

It's solution, yielding the eigenvalues λ_p and associated eigenvectors V_p of the matrix A , is therefore the same as in the principal components analysis, and thus the solved problem satisfies the problem of maximizing the discriminant criterion.

In the last equation, the number of non-zero eigenvalues of a square matrix A is equal to the rank of A . With $W^{-1}B$ playing the role of A , the number of non-zero eigenvalues depends on the rank of B , since the rank of the product of two matrices can not exceed the smaller of the two factor matrices' ranks, and W^{-1} (being nonsingular) must be of full rank p , while the rank of B is usually smaller than p . Thus it is possible to denote the rank of B by $r = \min (G-1, p)$.

From the fact that the eigenvalues λ_p are the values assumed by the discriminant criterion for linear combination using the elements of the corresponding eigenvectors P as combining weights, it is clear that the eigenvector

$V_1' = (v_{11}, v_{12}, \dots, v_{1p})$ provides a set of weights such that the transformed variable

$$Y_1 = v_{11}X_1 + v_{12}X_2 + \dots + v_{1p}X_p \quad (23)$$

has the largest discriminant-criterion, λ , achievable by any linear combination of the p independent variables.

What are the properties of the remaining eigenvectors, v_2, v_3, \dots, v_p ? The second discriminant function $Y_2 = v_{21}X_1 + v_{22}X_2 + \dots + v_{2p}X_p$ whose weights are the elements of the eigenvector v_2 associated with the second largest eigenvalue λ_2 of $W^{-1}B$ has the largest discriminant-criterion among those linear combinations of the X_i that are uncorrelated with the first discriminant function in the total sample observation. Its proof is analogous of that of principal components analysis. Each discriminant function has a relative (or conditional) maximum value for its discriminant criterion. Therefore, it needs nonly to show that Y_2 is uncorrelated with Y_1 . Noting that this correlation is proportional to $v_1'Tv_2$ (where $T = W + B$), we have to prove that $v_1'Tv_2 = 0$.

$$(B - \lambda_i W)v_i = 0 \quad \text{for each } i \quad (24)$$

hence,

$$Bv_1 = \lambda_1 Wv_1 \quad \text{and} \quad Bv_2 = \lambda_2 Wv_2$$

premultiplying these equations by v_2' and v_1' respectively,

$$\begin{aligned} v_2' B v_1 &= \lambda_1 v_2' W v_1 \\ v_1' B v_2 &= \lambda_2 v_1' W v_2 \end{aligned} \quad (25)$$

taking the transpose of both sides of the first equation (B and W are symmetric)

$$v_1' B v_2 = \lambda_1 v_1' W v_2$$

thus

$$\lambda_1 v_1' W v_2 = \lambda_2 v_1' W v_2$$

$$(\lambda_1 - \lambda_2) v_1' W v_2 = 0$$

since $\lambda_1 \neq \lambda_2$, $v_1' W v_2 = 0$

therefore, $v_1' W v_2 = 0$ which means V_1 and V_2 are uncorrelated, and Y_2 has this property: its discriminant-criterion value, λ_2 , is the largest achievable by any linear combination of X's that is uncorrelated (in the total sample) with Y_1 . Similarly

$$Y_3 = v_{31} X_1 + v_{32} X_2 + \dots + v_{3p} X_p \quad (36)$$

has the largest possible discriminant-criterion value (λ_3) among all linear combinations of the X's that are uncorrelated with both Y_1 and Y_2 ; and so on until Y_r using the

elements of V_r as weights, has the largest possible discriminant-criterion value among linear combinations that are uncorrelated with all the preceding linear combinations Y_1, Y_2, \dots, Y_{r-1} . The linear combinations Y_1, Y_2, \dots, Y_r are called the first, second, ..., r th (linear) discriminant functions for optimally differentiating among the g given groups.

The situation here is reminiscent of principal components analysis. There, the dimension corresponding to the first component had maximum variance; the second-component dimension had maximum variance among those uncorrelated with the first; and so on. In discriminant analysis, the ratio of between-to within-groups sums-of-squares merely takes the place of variance as the criterion in determining the successive dimensions. However, an important difference between the dimensions identified in discriminant analysis and those in component analysis is that the former are generally not mutually orthogonal in the test space, even though they are uncorrelated. That is, the axis representing the discriminant functions are not a subset of axes obtainable by rigid rotation of the original system of p axes; the discriminant rotation is an oblique rotation.

Just as in the principal components analysis, the dimensions represented by the discriminant functions may be interpreted meaningfully. Even if they are not, it may be possible to achieve parsimony by reducing the dimensionality of the space needed to describe group differences. In

seeking to interpret the discriminant functions, the goal is to determine which of the original p variables contribute most to each function. For this purpose, comparison of the relative magnitudes of the combining weights as given by the elements of each eigenvector of $W^{-1}B$ is inappropriate because these are weights to be applied to the variables in raw-score scales, and are hence affected by the particular unit used for each variable.

To eliminate the spurious effects of units of measurement on the magnitudes of combining weights, standardized variables should be used.

The relative magnitudes of these standardized weights may be assessed by multiplying each raw-score weight by the standard deviation of the corresponding variable as computed from the within-groups SSCP (Sum of Squares, Cross product) matrix. This amounts to multiplying each element of a given eigenvector V_m by the square root of the corresponding diagonal element of W . Thus, for each m , define

$$v_{mi}^* = w_{ii} v_{mi} \quad i = 1, 2, \dots, p \quad (27)$$

as the standardized discriminant weights. The relative contribution of the i th variable to the m th discriminant function may then be gauged by the magnitude of v_{mi}^* in comparison with the other weights v_{mj}^* .

Up to this point, it has been shown that the dimensionality of the discriminant space is equal to the number of

nonzero eigenvalues of $W^{-1}B$, which is the smaller of the two numbers, $G-1$ and p . It may often happen, that the number of significant discriminant dimensions may be even smaller. That is, not all of the discriminant function may represent dimensions along which statistically significant group differences occur.

C. SIGNIFICANCE TEST IN DISCRIMINANT ANALYSIS

A basic quantity in testing the significance of the overall difference among several group centroids (mean vectors) the ratio of the determinants of the within-groups and the total SSCP matrices, known as Wilks' Λ criterion.

$$\Lambda = \frac{|W|}{|T|} \quad (28)$$

Motivation for use of this equation may be seen as follows:

$$\begin{aligned} \frac{1}{\Lambda} &= \frac{|T|}{|W|} = |W^{-1}T| \\ &= |W^{-1}(W + B)| \\ &= (1 + \lambda)(1 + \lambda_2); \dots, (1 + \lambda_r) \end{aligned} \quad (29)$$

where $\lambda_1, \lambda_2, \dots, \lambda_r$ are the nonzero eigenvalues of $W^{-1}B$. Consequently, Bartlett's V statistic for testing the significance of an observed value can be expressed as

$$\begin{aligned}
V &= - [N - 1 - (p + G)/2] \ln \Lambda \\
&= [N - 1 - (p + G)/2] \ln [(1 + \lambda_1)(1 + \lambda_r)] \quad (30) \\
&= [N - 1 - (p + G)/2] \sum_{m=1}^r \ln(1 + \lambda_m)
\end{aligned}$$

This statistic is distributed approximately chi-square with $p(G-1)$ degrees of freedom.

Because of the uncorrelatedness of the successive discriminant functions, the successive terms $\ln(1 + \lambda_m)$ in the last expression above are statistically independent (assuming multivariate normality of the original p variables). As a result, the additive components of V are each approximately distributed as a chi-square variate. More specifically, the m th component,

$$V_m = [N - 1 - (p + G)/2] \ln (1 + \lambda_m) \quad (31)$$

is approximately chi-square with $p + G - 2m$ degrees of freedom. It may be readily verified that the sum of the number of degree of freedom (n.d.f) of the r components, that is, $(p + G - 2) + (p + G - 4) + \dots + (p + G - 2r)$, is equal to $p(G - 1)$ regardless of whether $r = G - 1$ or p .

Consequently, when we cumulatively subtract V_1, V_2 , and so on from V , the remainder each time is also a chi-square variate; and these successive remainders become appropriate statistics for testing whether the residual discrimination after removing the first discriminant

function, the first and second discriminant function, and so forth, is statistically significant. The successive test statistics and their n.d.f.'s may be summarized as follows:

Residual After Removing	Approximate - Statistic	n.d.f.
First discriminant Function	$V - V_1$	$p(G-1) - (p+G-2)$ $= (p-1)(G-2)$
First 2 discriminant Function	$V - V_1 - V_2$	$(p-1)(G-2) - (p+G-4)$ $= (p-2)(G-3)$
First 3 discriminant Function	$V - V_1 - V_2 - V_3$	$(p-2)(G-3) - (p+G-6)$ $= (p-3)(G-4)$
⋮	⋮	⋮
First s discriminant Function	$V - V_1 - V_2 - V_3 \dots - V_s$	$(p-s)(G-(s+1))$

As soon as the residual, after removing the first s discriminant functions becomes smaller than the prescribed percentile point (that is, the $100(1 - \alpha)$ th percentile) of the appropriate chi-square distribution, we may conclude that only the first s discriminant functions are significant at that α level. If the number of significant discriminant functions thus found is smaller than r (as will often be the case), we will have effected a further reduction in the dimensionality of the space required to describe the differences among the G groups from which

our sample groups were drawn. The remaining $r-s$ dimensions may be regarded as immaterial for population differentiation, since our sample differences along these dimensions can be attributed to sampling error.

IV. CLUSTER ANALYSIS

A. ORIGIN AND THEORY

Clustering is the grouping of similar objects. The principal functions of clustering are to name, to display, to summarize, to predict, and to aid in interpretation of data with many dimensions. Clustering techniques were first developed in the field of biological taxonomy. It is one of several methodologies included in the broader category called classification.

The cluster analysis problem is the last step we consider in the progression of category sorting problems. While in discriminant analysis some part of the structure is known and missing information is estimated from labeled samples, the operational objectives of clustering is to classify new observations, that is, recognize them as members of one category or another. In cluster analysis little or nothing is known about the category structure. All that is available is a collection of observations whose category membership are known. We seek to discover a category structure which fits the observations. The problem may be stated as one of finding the "natural groups", which means to sort the observations into groups such that the degree of "natural association" is high among members of the same group and low between members of different groups.

Cluster analysis techniques have been applied in many fields of study. The literature is both voluminous and diverse, the terminology differing from one field to another. "Numerical taxonomy" is frequently substituted for cluster analysis among biologists, botanists, and ecologists, while some social scientists may refer "typology". Other frequently encountered terms are pattern recognition and partitioning. While discriminant analysis has been studied by statisticians for nearly 45 years, cluster analysis has only recently come to statistical notice. Any method which partition a set of objects into subsets on the basis of measurements taken on every object qualifies as a clustering method.

Most of the well known clustering techniques fall into one of two main categories: (1) hierachical and (2) non-hierachical (partitioning). The former is one in which every cluster obtained at any stage is a merger of clusters at previous stages. The nonhierachial procedures however form new clusters by lumping and splitting old ones. We consider both categories shortly.

In a geometric sense, every observation may be viewed as a point in p -dimensional Euclidean space. This swarm of data points may contain dense regions or "clouds" of data points which are separable from other regions containing a low density of points. These denser regions constitute what are known as clusters. In one and two dimensional cases, it is easy to visualize and to detect the clusters from scatter

plots, assuming that the clusters exist. In higher dimensions, clustering becomes extremely difficult without the aid of a computer.

Mathematical clustering techniques usually require a measure of similarity to be defined for every pairwise combination of the entities to be clustered. In order to solve the cluster problem, it is desirable to define the terms "similarity" and "difference" in a quantitative fashion. A researcher would assign two observations to the same group if the distance between them is sufficiently small, or to different clusters if this distance is sufficiently large.

At this point, two questions may be brought on. The first one is "how do we measure the distance between the observations?" and the second one is "how small is small enough?" and how large is large enough? These will be discussed in the following sections.

B. MEASURES OF DISTANCE

1. General

Let E_p be a symbolic representation for a measurement in p -dimensional space and let X, Y , and Z be any of these points in E_p . Then any nonnegative real-valued function $D(X, Y)$ satisfying the following conditions qualifies as a distance function (or metric).

1. $D(X, Y) = 0$ if and only if $X = Y$
2. $D(X, Y) \geq 0$ for all X and Y in E_p

$$3. \quad D(X,Y) = D(Y,X)$$

$$4. \quad D(X,Y) \leq D(X,Z) + D(Y,Z)$$

Many clustering algorithm assume such distances given and set about constructing clusters of objects within which the distances are small. The choice of distance function is no less important than the choice of variables to be used in the study. A serious difficulty in choosing a distance lies in the fact that a clustering structure is more primitive than a distance function and that knowledge of clusters changes the choice of distance function. Thus a variable that distinguishes well between two established clusters should be given more weight in computing distances than a "junk" variable that distinguishes badly.

2. Euclidean Distance

The Euclidean distance between the I-th and K-th observations of a data matrix X is defined as

$$D(I,K) = \left[\sum_{1 \leq J \leq p} \{X(I,J) - X(K,J)\}^2 \right]^{1/2} \quad (32)$$

where J is J-th variable. In one, two, or three dimensional space, this is just a "straight line" distance between the vectors corresponding to the I-th and K-th observations. When the variables are measured in different units, it is necessary to prescale the variables to make their values comparable or, equivalently, to compute a weighted Euclidean distance.

$$D(I,K) = \left[\sum_{1 \leq J \leq p} W(J) (X(I,J) - (X(K,J)))^2 \right]^{1/2} \quad (33)$$

This form of distance is not necessary if all variables are measured on the same scale. However, even in this case, weights might be used to increase or decrease the importance of same variable. Various weighting schemes have been utilized in practice. One common weighting scheme lets $W(J)$ be the reciprocal of the variance of variable J .

A general class of squared distance functions is provided by utilizing positive definite quadratic forms. Specifically, if p represents a p -dimensional observation to be assigned to one of s groups, then to measure the squared distance between the observation β and the centroid (mean vector) of the i -th group one may consider the function

$$D_i = (\beta - \bar{x}_{i.})^T M (\beta - \bar{x}_{i.})$$

where M is a positive definite matrix to ensure that $D_i \geq 0$. Different distance functions are represented by different choices of the matrix M . When $M = I$ (the identity matrix) the resulting metric is the standard Euclidean distance. Distances with the Euclidean metric are shown in Figure 1a. The variance within the data may make the unweighted Euclidean metric inappropriate. As shown on the Figure 1b, where X has a larger variance than Y ,

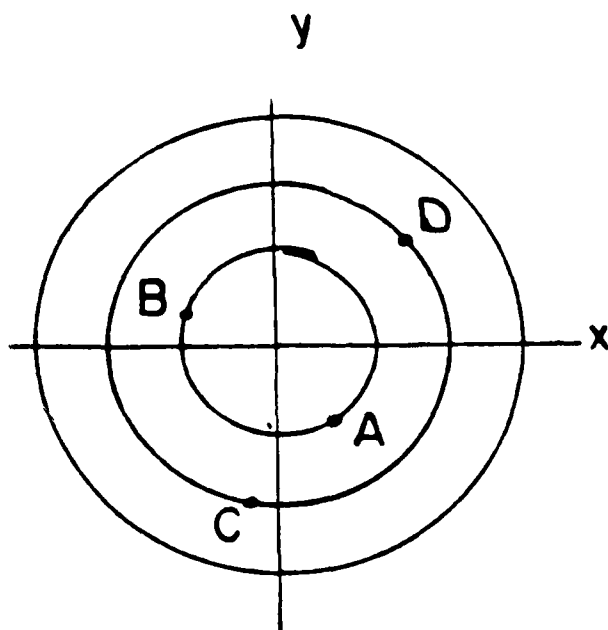
one may wish to weight a deviation in the X direction less than an equal deviation in the Y direction. This is a weighted Euclidean distance function which makes point A and B equidistance from the origin. In this case, the matrix M is diagonal elements which are the reciprocals of the variances of the different variables.

Extending this idea further, it may be possible to consider the covariance among variables as well. Figure 1c shows how the axis may be rotated so that the major axis is oriented in a direction of reflecting the positive correlation between X and Y. Again, points on the same ellipse are considered equidistance from the origin. The matrix M in this case is the inverse of the covariance matrix.

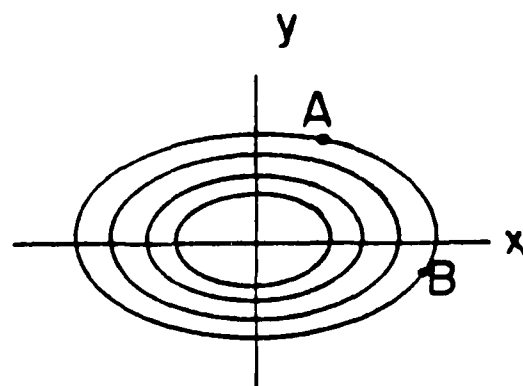
Further extension of this concept will explain some sort of generalized distance function. If C_i represents the covariance matrix of the i th cluster then the distance function

$$D_i = (\beta - \bar{x}_{i.})^T C_i^{-1} (\beta - \bar{x}_{i.})$$

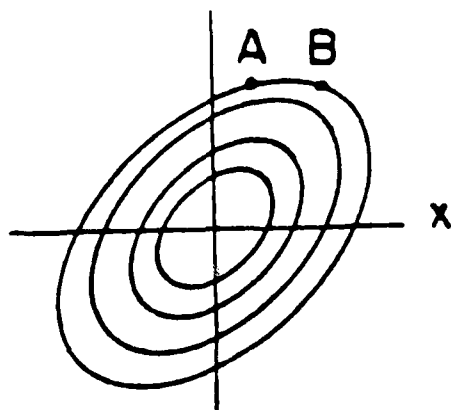
uses the appropriate covariance structure when determining the distance to a particular cluster centroid. Since C_i changes to reflect the dispersion internal to each particular cluster, the use of this metric exploits differences in the dispersion characteristics of the different groups. As shown on Figure 1d, not how a new observation (denoted by u) is



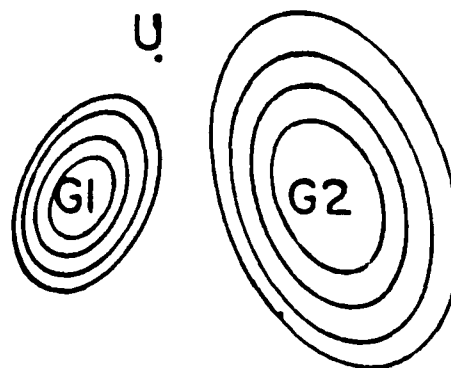
1a. Euclidean measure of squared distance.



1b. Measure of squared distance with different weights for variables.



1c. Generalized squared distance measure.



1d. Classification when within-group dispersions are different.

Figure 1. Euclidean Distance

closer to the centroid of group one (G1) in terms of Euclidean distance but is more likely to be assigned to group two (G2) when using the C_i matrix.

3. Mahalanobis Distance

Another choice for the M matrix in equation (1) is p^{-1} where P represents the pooled within groups covariance matrix of all the clusters.

$$P = \frac{1}{\sum_{i=1}^G (n_i - 1)} W \quad (34)$$

where

$$W = \sum_{k=1}^G W_k$$

This distance is the well known Mahalanobis distance. Note that P does not change from group to group. To ensure the non-singularity of P it must be true that $p \leq (N - G)$, where N represents the total number of observations over all groups. Rewriting the distance,

$$D_i = (\beta - \bar{x}_{i.})^T W^{-1} (\beta - \bar{x}_{i.}) \quad (35)$$

defines a distance between mean vectors β and $\bar{x}_{i.}$ and common covariance matrix W . The Mahalanobis distance function adjusts for both scale of measurement of the variables and covariation among the variables. Use of this

metric is equivalent to computing distances on variables transformed to their principal components. This metric is invariant under any nonsingular transformation of original variables. For consider the transformation

$$Y = BX \quad (36)$$

and let $D(Y_i, Y_j)$ represent Mahalanobis distance between Y_i and Y_j .

$$\begin{aligned} D(Y_i, Y_j) &= (Y_i - Y_j)^T P_Y^{-1} (Y_i - Y_j) \\ &= (BX_i - BX_j)^T P_Y^{-1} (BX_i - BX_j) \\ &= (X_i - X_j)^T B^T P_Y^{-1} B (X_i - X_j) \\ &= (X_i - X_j)^T B^T (B P_X B^T)^{-1} B (X_i - X_j) \\ &= (X_i - X_j)^T P_X^{-1} (X_i - X_j) \\ &= D(X_i, X_j) \end{aligned}$$

Some other common metrics are listed below:

1. L_1 norm (City Block)

$$D(X_i, X_j) = \sum_{k=1}^P |X_{ki} - X_{kj}|$$

2. L_p norm (Minkowsky Metrics)

$$D(X_i, X_j) = \left(\sum_{k=1}^P |X_{ki} - X_{kj}|^p \right)^{1/p}$$

3. Uniform norm

$$D(X_i, X_j) = \text{Superemum}_{k=1, 2, \dots, p} \{ |X_{ki} - X_{kj}| \}$$

C. HIERARCHICAL CLUSTERING

1. General

The previously discussed distance measures may be used to construct a similarity matrix describing the length of all pairwise relationships among the entities (variables or data units) in the data set. The methods of hierarchical cluster analysis operate on this similarity matrix to construct a tree depicting specified relationships among the entities. As shown on Figure 2, the branches on the left each represent one entity while the root represents the entire collection of entities. Moving down the tree from the branches toward the root depicts increasing aggregation of the entities into clusters. Hierarchical clustering methods which build a tree from branches to root often are called agglomerative methods.

Once a tree is constructed for N entities, the analyst may choose from as many as N sets of clusters. These clusters are nested. From the agglomerative view, when two entities are merged they are joined together permanently and considered as one entity for later merges; from the divisive view, when a group of entities is split into two parts, the parts are separated permanently and may be treated independently for the remainder of the analysis.

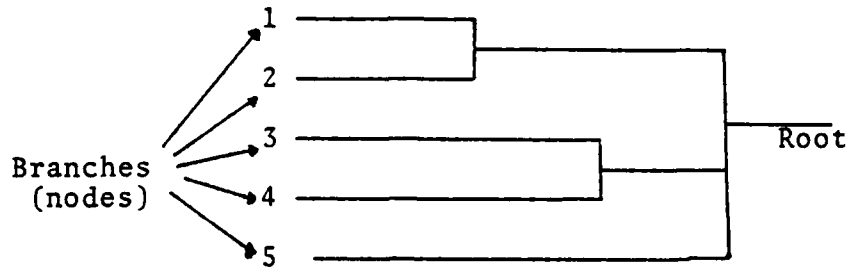


Figure 2. Tree for Hierarchical Clustering

Herein lie both the strength and weakness of hierarchical methods: by taking early decisions as permanent, the number of possibilities that need be examined is reduced greatly as compared with complete enumeration; but this same convention precludes discovering early mistakes or capitalizing on later opportunities.

There are three major hierarchical clustering concepts:

1. Linkage Methods
2. Centroid Methods
3. Error sum of squares or variance methods.

All of these methods are suitable for clustering data units. However, only the linkage methods are considered in this research.

2. The General Agglomerative Procedure

Let s_{ij} be the similarity between entities i and j as defined by one of the distance measures previously discussed. Assuming that the similarity is symmetric, the complete schedule of similarities for all $\binom{N}{2} = \frac{1}{2}N(N - 1)$

possible pairwise combinations of entities may be arrayed in a lower triangular similarity matrix as in Figure 3. The s_{ij} entries are nonnegative. This limitation is of consequence only for correlation and the cosine of the angle between vectors; the distinction between positive and negative association cannot be utilized in these clustering methods.

$$S = \begin{array}{c|cccc} & s_{21} & & & \\ & s_{31} & s_{32} & & \\ & s_{41} & s_{42} & s_{43} & \\ & . & . & . & . \\ & . & . & . & . \\ & s_{n1} & s_{n2} & s_{n3} & \cdots s_{n(n-1)} \end{array}$$

Figure 3. Lower Triangular Similarity Matrix

A simple remedy is to use the absolute value or the square of the measure if it can assume negative values. Once the matrix is defined, the process of clustering entities is almost trivially simple. The general procedure for agglomerative clustering on a data matrix is as follows:

- (1) Begin with n clusters each consisting of exactly one entity. Let the clusters are labeled with the numbers 1 through N .

- (2) Search the similarity matrix for the most similar pair of clusters. Let the chosen clusters be labeled p and q and let their associated similarity be s_{pq} , $p > q$.
- (3) Reduce the number of clusters by 1 thorough merger of clusters p and q . Label the product of the merger q and update the similarity matrix entities in order to reflect the revised similarities between cluster q and all other existing clusters. Delete the row and column of S pertaining to cluster p .
- (4) Perform steps 2 and 3 a total of $N-1$ times (at which point all entities will be one cluster). At each step record the identity of the clusters which are merged and the value of similarity between them in order to have a complete record of the results.

Different agglomerative methods are implemented by varying the procedures used for defining the most similar pair at step 2 and for updating the revised similarity matrix at step 3. The similarity matrix is a given array of numbers. The numerical execution of the clustering procedures is completely independent of how the similarity values were generated or whether the entities to be clustered are variables or data units. However, it is necessary to make a direct distinction between distance-like

measures (the smallest values correspond to the most similar pairs) and correlation-like measures (the largest values correspond to the most similar pairs); the essential difference is whether the search for the most similar pair involves seeking the minimum or maximum entry in the similarity matrix.

3. Single Linkage

The method of single-linkage cluster analysis is the simplest of all hierarchical techniques. At each stage, after clusters p and q have been merged, the similarity between the cluster (labeled t) and some other r is determined as follows:

1. If s_{ij} is the distance-line measure

$$s_{tr} = \min (s_{pr}, s_{qr}) \quad (37)$$

The quantity s_{tr} is the distance between the two closest members of clusters t and r . If clusters t and r were to be merged, then for any entity in the resulting cluster the distance to its nearest neighbor would be at most s_{tr} .

2. If s_{ij} is a correlation-like measure

$$s_{tr} = \max (s_{pr}, s_{qr}) \quad (38)$$

The quantity s_{tr} is the similarity between the two most similar entities in clusters t and r . If clusters t

and r were to be merged, then for any entity in the resulting cluster there would be at least one other entity in the same cluster such that the pair would have a similarity at least as large as s_{tr} .

The method is known as single linkage because clusters are joined at each stage by the single shortest or strongest link between them. Since the updating process involves choosing only the minimum or maximum single-linkage clustering is invariant to any transformation which leaves the ordering of the similarities unchanged; that is, any monotonic transformation.

4. Complete Linkage

The complete-linkage method is related to the single-linkage method and is no more difficult to execute. At each stage, after clusters p and q have been merged, the similarity between the new cluster (labeled t) and some other cluster r is determined as follows:

1. If s_{ij} is distance-like measure

$$s_{tr} = \max (s_{pr}, s_{qr}) \quad (39)$$

The quantity s_{tr} is the distance between the most distant members of clusters t and r . If clusters t and r were merged, then every entity in the resulting cluster would be no farther than s_{tr} from every other entity in the cluster. The value of s_{tr} is the diameter of the smallest sphere which can enclose the cluster resulting from the merger of clusters t and r .

2. If s_{ij} is a correlation-like measure

$$s_{tr} = \min (s_{pr}, s_{qr}) \quad (40)$$

The quantify s_{tr} is the similarity between the two most dissimilar entities in clusters t and r . If clusters t and r were to be merged, then every entity in the resulting cluster would have a similarity of at least s_{tr} with every other entity in the cluster.

The method is called complete linkage because all entities in a cluster are linked to each other at some maximum distance or minimum similarity. Such a cluster is called a "maximally connected subgraph" in graph theory. In contrast to the single-linkage method, interpretation of the clusters can be made only in terms of the relationships within individual clusters; there is no particularly useful interpretation involving the differences between clusters. Like the single-linkage method, complete-linkage cluster analysis is invariant to monotonic transformations of the similarity measure. Johnson (1967) discusses this property in both single and complete linkage methods.

D. NONHIERARCHICAL CLUSTERING

Nonhierarchical clustering methods are designed to cluster data units into a single classification of g clusters, where g either is specified a priori or is determined as a part of the clustering method. The central idea in most of these methods is to choose some initial

partition of the data units and then alter cluster memberships so as to obtain a better partition. The various algorithms which have been proposed differ as to what constitutes a "better partition" and what methods may be used for achieving improvements.

The broad concept for these methods is very similar to that underlying the steepest descent algorithms used for unconstrained optimization in nonlinear programming. Such algorithms begin with an initial point and then converge to a local optimum, moving one step at a time, the value of the objective function improving at each step.

The methods of nonhierarchical clustering typically may be used with much larger problems than the hierarchical methods because it is not necessary to calculate and store the similarity matrix; it is not even necessary to store the data set. In general, the data units are processed serially and can be read from tape or disk as needed. This characteristic makes it possible, at least in principle, to cluster arbitrary large collections of data units.

In this research, we consider only the partitioning method known as "K-MEANS" which was developed by MacQueen (15). He used the term "K-MEANS" to denote the process of assigning each data unit to that cluster (of k clusters) with the nearest centroid (mean vector). The cluster centroid changes with each transfer of an observation.

The decomposition of the total scatter matrix into within and between groups matrices suggests possible

optimality criteria to be used in a clustering algorithm. One would like the within-groups scatter to be small relative to the between-groups scatter. Various trial clusterings could be formed using the W and B matrices as a basis for the optimality criteria which determine the best clustering. A possible choice for a criterion is to minimize trace W over all partitions into g groups. Since T is constant over all partitions, minimizing trace W is equivalent to maximizing traces B since

$$\text{trace } T = \text{trace } W + \text{trace } B \quad (41)$$

Although trace W is invariant under an orthogonal transformation, it is not invariant under other non-singular linear transformations.

McRae (16) points out that trace W equals the total within group sum of squares, hence the "minimum variance partition" cluster solution is found by minimizing trace W .

Considerable study has been developed to alternative criteria such as those based on multivariate statistical analysis techniques, especially the methods of linear discriminant analysis and multivariate analysis of variance. Assuming the p variables are not linearly dependent, then as long as $p = N - g$, W is positive definite symmetric and so is W^{-1} . Attempts to make B and W as different as possible lead one to solving the determinantal equation:

$$|B - \lambda W| = 0 \quad (42)$$

The solutions λ_i are the eigenvalues of the matrix $W^{-1}B$ as in discriminant analysis. There are t non-zero eigenvalues, where t is the minimum of p and $g-1$. This is a consequence of the fact that, if g is less than p , the g group means are considered in a $(g-1)$ -dimensional hyperplane. When $g = 2$ the analysis is equivalent to two-group discriminant analysis. Linear discriminant analysis would take the vectors originally described in p -dimensional coordinate system and transform the basis to a t -dimensional system. Maximizing the largest of these eigenvalues is a criterion suggested by S.N. Roy and maximizing the trace of $W^{-1}B$, however is a criterion suggested by Hotelling. In both cases, large values for these statistics are sought in clustering algorithms since large values indicate large differences among (between) groups. Minimizing the ratio of determinants $|W| \div |T|$ is a criterion widely known as Wilks' lambda discussed in the discriminant analysis. Since T is the same for all partitions, this criterion is equivalent to minimizing determinant W . Both trace $W^{-1}B$ and $|T| \div |W|$ may be expressed in terms of the eigenvalues of $W^{-1}B$.

$$\frac{|T|}{|W|} = \prod_{i=1}^t (1 + \lambda_i) \quad (43)$$

$$\text{trace } W^{-1}B = \sum_{i=1}^t \lambda_i \quad (44)$$

where $t = \min(p, g-1)$. Therefore minimizing $\det W$ is equivalent to maximizing $\pi(1 + \lambda_i)$.

Friedman and Rubin (6) describe the advantages of the various criteria. Those based on multivariate statistical considerations (all but trace W) are invariant under changes in scale for variables (non-singular linear transformation). In fact, they are the only invariants for W and B under such transformations. In addition, the multivariate criteria may take into account covariation among the variables.

V. ANALYSIS OF MULTIVARIATE UTILITY DATA

To illustrate hierarchical clustering we applied the technique described in the previous chapter to partition a set of twenty six attributes of a close-air support weapon system into a smaller collection of "superattributes". As part of an effort to evaluate the military utility of a proposed alternative U.S. Marine Corps air support rada system, AN-TPQ/27. Barr and Richards (4) extracted 26 attributes of the TPQ-27 and a baseline system, the AN-TPQ/10, and then had members of the Operational Test and Evaluation Team assess the utility of the TPQ/27 relative to that of the TPQ/10. In order that the additive model used to combine unidimensional relative utilities into a system relative utility be justifiable, it is necessary that the utilities satisfy certain independence properties described in Keeney and Raiffa (12).

Because those independence properties are very difficult for decision makers to verify for complex alternatives like the weapon systems under study, Professors Barr and Richards attempted instead to work with the attributes to try to generate a new collection which would likely satisfy, at least approximately, the conditions required to justify the additive model.

The original collection of 26 attributes is as follows:

1. Portability
2. Durability
3. Time to Set Up
4. Time to Take Down
5. Ease of Assigning Aircraft to Targets
6. Number of Aircraft Controlled
7. Number of Targets
8. Communications
9. Mission Flexibility
10. ASRT Survivability
11. Time to Locate and Acquire Aircraft
12. Accuracy of Tracking
13. Accuracy of Delivery
14. Range
15. Aircraft Vulnerability
16. Aircraft Attack Throughout
17. Base of Adjustment and Evaluation of Results
18. Accuracy of Feedback
19. Ease of Operation
20. Man-Machine Compatibility
21. Training Requirements
22. Reliability
23. Maintainability
24. Supportability
25. Availability
26. Documentation

where a_i represents an attribute i and

$$I(x_{ij}, x_{kj}) = \begin{cases} 1 & \text{if } x_{ij} = x_{kj} \\ 0 & \text{if } x_{ij} \neq x_{kj} \end{cases}$$

It is easy to verify that D is a metric as defined in Chapter IV. Since we will actually work with a similarity measure in the hierarchical cluster procedure, we define the similarity between two attributes a_i and a_k as

$$S(a_i, a_k) = \sum_{j=1}^{12} I(x_{ij}, x_{kj}) \quad (46)$$

One can see from this definition that the similarity between two attributes a_i and a_k is simply the number of team members who placed attributes a_i and a_k in the same partition. For example,

$$\begin{aligned} S(a_1, a_2) &= 0 + 1 + 0 + 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0 \\ &\quad + 1 + 0 = 4 \end{aligned}$$

Either S or D can be used in the computer program shown in Appendix A for hierarchical clustering. One need only indicate whether he wants a correlation-like (larger values imply more similar) measure or a distance-like measure (smaller values imply more similar). We selected to use the former method. The similarity matrix extracted from

Table II. Data Matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	6	2	1	3	2	4	3	1	3	1	1	1
2	1	2	2	3	1	3	2	1	1	5	1	2
3	6	1	1	5	2	4	3	1	6	1	2	1
4	6	1	1	5	2	5	3	1	6	1	2	1
5	2	5	3	1	3	1	1	2	2	2	3	3
6	2	7	4	1	4	1	1	2	4	2	3	3
7	2	7	4	1	4	1	1	2	4	2	3	3
8	3	5	4	7	5	6	1	3	1	3	3	6
9	2	5	4	1	3	1	1	2	4	2	3	3
10	9	6	5	6	5	7	8	3	7	3	2	4
11	2	8	6	4	3	2	1	4	4	2	3	3
12	3	8	7	4	4	2	6	5	4	6	4	3
13	3	8	7	4	4	2	6	5	4	6	4	3
14	3	8	4	4	4	2	6	5	4	2	4	3
15	7	6	5	6	5	7	7	3	7	3	7	4
16	8	8	6	1	4	1	7	4	4	2	7	3
17	8	8	8	4	3	2	5	6	5	4	5	3
18	4	8	8	4	6	2	5	6	5	6	5	3
19	4	5	3	1	3	1	1	2	2	4	3	3
20	4	5	3	3	3	3	3	1	2	4	8	3
21	5	4	9	2	3	8	4	7	2	4	6	5
22	1	3	2	3	1	3	2	7	1	5	6	2
23	1	3	9	3	1	3	2	7	1	5	6	2
24	1	3	4	3	1	3	2	7	3	5	6	2
25	1	3	2	3	1	3	2	7	1	5	6	2
26	5	4	9	2	6	8	4	7	2	4	6	5

from the data is shown in Table V-3. We present only lower triangular elements since $S(a_i, a_i) = 12$ for all i and the matrix is symmetric; i.e., $S(a_i, a_k) = S(a_k, a_i)$. Zero values are not written.

The results from the hierarchical clustering are shown in Figure 4. The numbers printed along the left hand margin refer to the attribute numbers. As you proceed to the right through the tree you will observe numbers greater than 26. These correspond to the clusterings that takes place from one step to the next. For example, the number 27 shown at the juncture of 25 and 22 means that the first attribute clustered together should be 25 and 22 (this is the most similar pair). This combination is then considered as a new attribute which is later combined with the attribute 30 (itself a combination of 23 and 24) to form the attribute 31. This is later combined with attribute 2 to form attribute 40, etc.

As discussed in Chapter IV a decision has to be made as to how many clusters (superattributes) are desired. All hierarchical methods will continue clustering until there is a single cluster. In order to decide on the number of clusters (and their composition) one need only image drawing a vertical line through the tree at various places. Each intersection of the tree with the vertical line results in a cluster. For example, the vertical line at the point A results in the 6 clusters shown in Table V-4.

It is clear from observing the above collection that some of the attributes are highly correlated and nonredundant. If one tries to assign an importance weights to each attributes separately, there is a distinct likelihood that some of the overlapping strongly into related attributes might effectively be double or triple weighted or more producing biased result. It is an effort to prevent this from happening, Barr and Richards asked the utility assessment team to partition the 26 attributes into a smaller collection in such a way that attributes within a group are similiar and attributes in different groups are unrelated the sense that utility assessments for attributes in one group do not depend on the amounts of attributes in any other group.

The total number of groups was not prespecified. Instead, each team member was allowed to partition the 26 attributes into any number of groups. The resulting multivariate data array is shown in Table V-2. An element x_{ij} is the number of the group into each team member j put attribute i .

Let us define a distance measure for this data array as follows:

$$D(a_i, a_k) = \sum_{j=1}^{12} (1 - I(x_{ij}, x_{kj})) \quad (45)$$

Table III. Similarity Matrix for Superattribute Determination.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1																										
2	4																									
3	8	1																								
4	7	1	11																							
5																										
6					8																					
7					8	12																				
8		1			3	3	3																			
9					10	10	10	3																		
10			1	1				3																		
11					6	6	6	2	7																	
12					1	3	3	3	1	2		5														
13					1	3	3	3	1	2		5	12													
14					2	5	5	2	4			5	9	9												
15								3		9																
16					3	5	5		4	1	6	3	3	4	3											
17					2	1	1	1	2		5	4	4	3		2										
18					1	1	1	1	1		5	4	4	3		1	9									
19					9	5	5	2	7		3					2	2	1								
20	2	3	1	1	5	1	1	1	3		2						2	1	8							
21					1				1		1						2		3	3						
22	2	8						1												2	2					
23	2	7						1												2	3	11				
24	3	6						1												2	3	10	11			
25	2	1																		2	2	12	11	10		
26					1												1	1	2	2	9	3	4	4	3	

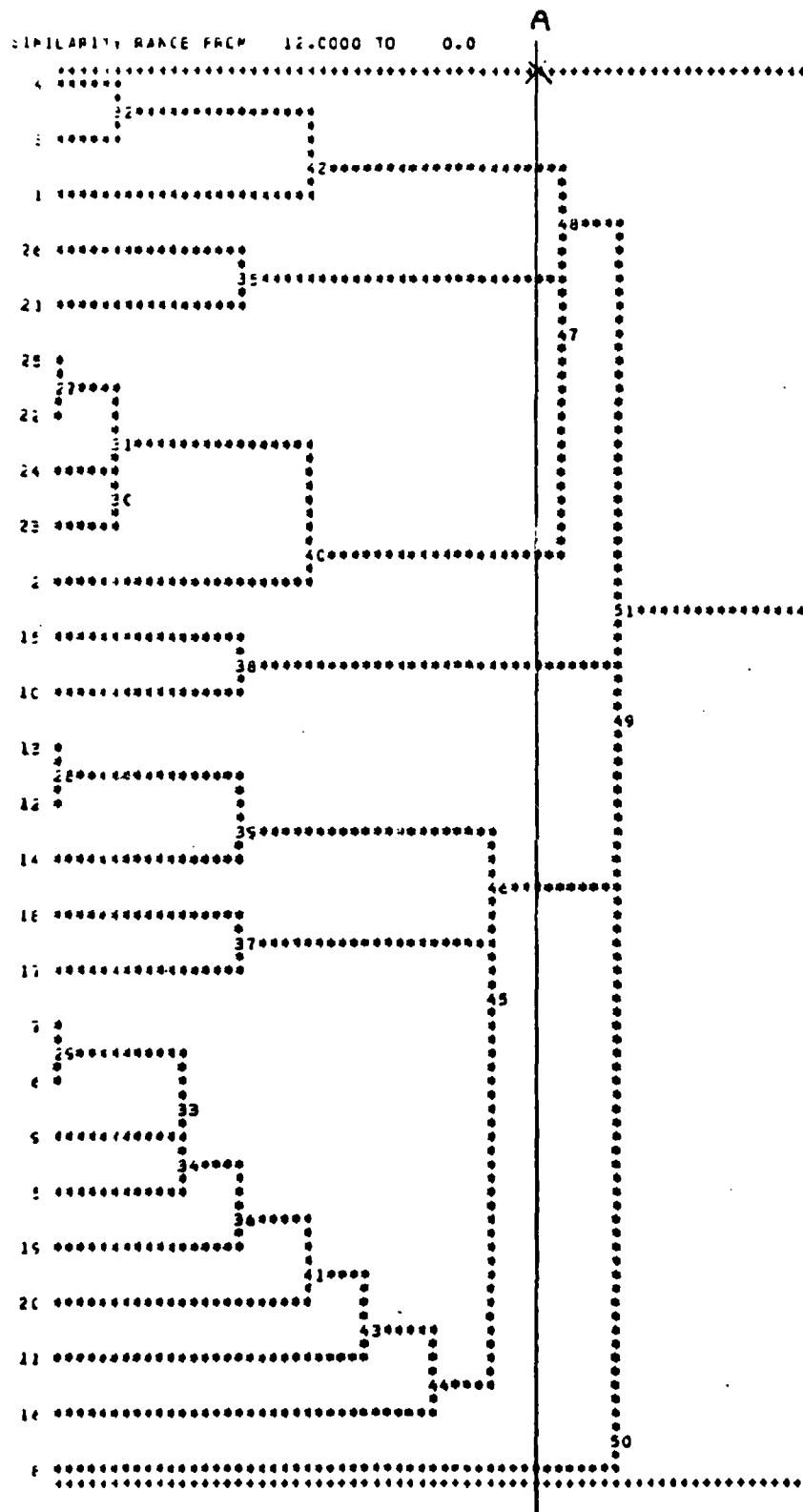


Figure 4. Tree for 26 Attributes

The superattributes used in the utility study are those shown in Table IV. A careful examination of the attributes which compare the clusters shows that the results so obtained are intuitively agreeable. The names supplied to the superattributes are somewhat natural descriptions of the clusters obtained.

The listing of the computer program and sample output are given in Appendix A.

Table IV. Superattributes

<u>Superattributes</u>	<u>Component Attributes</u>
Facility of movement	1. Portability 3. Time to set up 4. Time to take down
Facility of Use (precision)	5. Ease of assigning aircraft to targets 6. Number of aircraft controlled 7. Number of targets 9. Mission flexibility 11. Time to locate and acquire aircraft 16. Aircraft attack throughput 17. Ease of adjustment 18. Accuracy of feedback 19. Ease of operation 20. Man-machine compatibility 12. Accuracy of tracking 13. Accuracy of delivery 14. Range
Survivability	10. ASRT Survivability 15. Aircraft vulnerability
Learning	21. Training requirements 26. Documentation
Readiness	2. Durability 22. Reliability 23. Maintainability 24. Supportability 25. Availability
Communications	8. Communications

VI. ANALYSIS OF ARMY TANK DATA

A. DATA STRUCTURE

In order to illustrate the nonhierarchical clustering methodology, principal components analysis, and discriminant analysis data on Army tanks from eight different countries were taken from Jane's Book of Weapon Systems (1979-80). A total of twenty-four tanks were included in the data array with observation on each of 10 variables. The 10 variables are listed below:

1. Weight (ton)
2. Length (meter)
3. Width (meter)
4. Height (meter)
5. Road Speed (kilometer per hour)
6. Trench Crossing (meter)
7. Ground Pressure (Kg/cm^2)
8. Maximum Armament (rounds)
9. Ground Clearance (meter)
10. Power to Engine Ratio (BHP/ton)

The twenty-four tanks and the associated countries are shown below:

Identification Number	Type/Name	Country
11	T-62	
12	T-54	U.S.S.R.

Identification Number	Type/Name	Country
13	T-10	
14	ASU-85	
15	MK-5/Chieftain	
16	MK-3/Vickers	
17	MK-13/Centurion	U.K.
18	CVR(T)/Scorpion	
19	XM-1	
20	M60A2	
21	M60	U.S.A.
22	M48	
23	M47	
24	PZ61	
25	PZ68	SWITZER- LAND
26	STRV-103	
27	Ikv-91	SWEDEN
28	TYPE61	
29	TYPE74	JAPAN
30	Leopard 2	
31	Leopard	W. GER- MANY
32	TAM	
33	AMX 30	
34	AMX 13	FRENCH

We conjecture that a cluster analysis of the tank data will result in clusters corresponding to nationality since the nations may have different emphasis on the variables in the design of their tanks.

B. NONHIERARCHICAL CLUSTER ANALYSIS OF TANK DATA

1. The MIKCA Algorithm

The specific algorithm chosen for the nonhierarchical cluster analysis for the tank data is the MIKCA (Multivariate Iterative K-MEANS Clustering Algorithm) program written by Douglas J. McRae as a part of his doctoral dissertation at the University of North Carolina, Chapel Hill.

Reference to the flow chart in Figure 5 will aid the reader in following discussion of the algorithm. Inputs to program are the data matrix, an estimate for g (the number of clusters), and choice of criterion and distance functions.

In the first step, preliminary calculations are made, such as the variable means and standard deviations, as well as the cross product matrix T . The next step forms the initial cluster centers. Then each of the other observations is assigned to the nearest cluster. Euclidean distance is used for this initial phase, and the cluster centroids are recomputed after each observation is assigned to a group. The observations are considered in the same order as they were input. After all of them have been assigned to clusters, the criterion value is computed. This initial cluster-finding technique is referred to as a

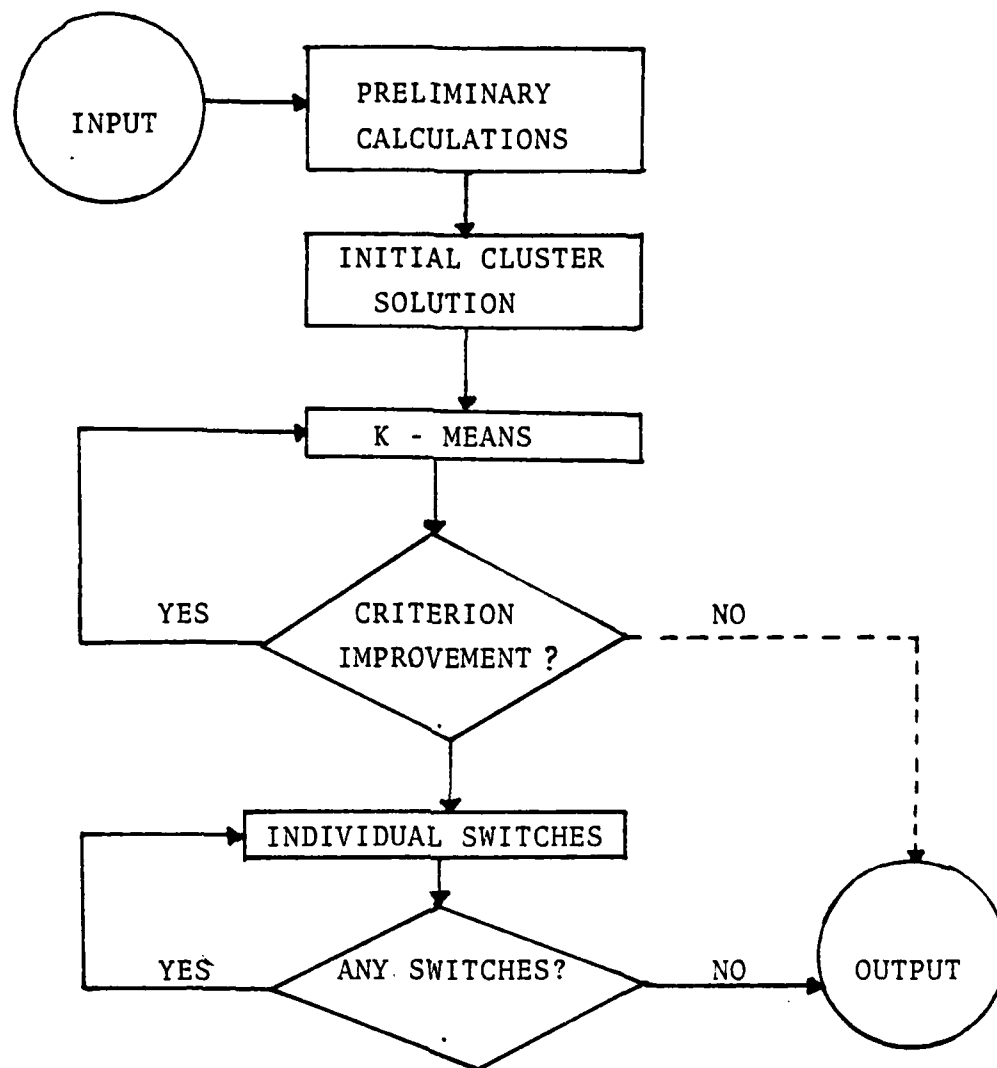


Figure 5. MIKCA Flow Chart

one-pass K-MEANS procedure. It is performed three times, and the solution which yields the best criterion value is chosen as the initial cluster solution.

After the initial solution has been found, the program advances to the iterative K-MEANS phase where the observations are again considered in the order in which they were input to the program. It is this phase where the user's choice of distance function is used. The distance from each observation to each cluster centroid is again computed, this time with the user's distance function, the assignment to the closest centroid being made and the centroid updated to reflect its new membership. After considering all n observations in this manner, the new criterion value is checked for possible improvement during the K-MEANS iteration. As long as the criterion value improves, the K-MEANS procedure is repeated; if the criterion fails to improve then the MIKCA algorithm goes to the next step, the individual switches section. Note the importance of the order of consideration of the observations. The order is important because the cluster means are recomputed after each observation is reassigned.

In the individual switches phase, consideration is given to moving each observation to every other cluster, the move being made if and only if an improvement in the value of the criterion results. An elaborate labelling procedure provides a unique order in which to consider each observation.

This procedure continues until a complete pass through the data is made with no changes in cluster membership.

The MIKCA algorithm provides the following options for distance and criterion functions.

Criterion

1. Minimum trace W
2. Minimum determinant W
3. Maximum largest order of $|B - \lambda W| = 0$
4. Maximum sum of roots of $|B - \lambda W| = 0$

Distance

1. Euclidean
2. Weighted Euclidean
3. Mahalanobis

A complete computer program is listed in Appendix B.

2. Cluster Results for Tank Data

For clustering of the tank data we selected the minimum trace W criterion and the weighted Euclidean distance function. The algorithm automatically provides weights for the weighted Euclidean distance function. The results of the clustering with four clusters are shown in Table V.

The conjecture of clustering by nationalities is supported by the results. The three Soviet tanks make up one cluster and the two British and four of the United States tanks were found to be similar. A third cluster consists of four tanks which are very lightweight. The

Table V. THE FINAL CLUSTER SOLUTION IS

CLUSTER 1									
SIZE = 11									
CENTER =	40.60	2.64	7.00	3.26	2.55	59.18			
THE OBSERVATIONS ARE	16	19	24	25	26	20.70			
	28	29	30	31	32				
	33								
CLUSTER 2									
SIZE = 6									
CENTER =	51.87	2.80	7.15	3.55	2.06	43.05			
THE OBSERVATIONS ARE	15	17	20	21	22	15.95			
	23								
CLUSTER 3									
SIZE = 4									
CENTER =	13.08	2.14	5.40	2.77	2.21	64.27			
THE OBSERVATIONS ARE	14	18	27	34	0.38	20.07			
CLUSTER 4									
SIZE = 3									
CENTER =	41.00	2.85	6.85	3.50	2.25	44.23			
THE OBSERVATIONS ARE	11	12	13	37.67	0.44	15.07			

final cluster consists of the rest of the tanks, including tanks of United States allies from West Germany, France, Sweden, Switzerland and Japan.

A natural question to ask after observing the results of a cluster analysis is what variables most strongly influence the clustering that was observed. A clue is provided by the composition of the cluster containing all of the lightweight tanks. This suggests that weight is an important distinguishing feature. This is examined in the principal components analysis and the discriminant analysis in the next two sections.

C. PRINCIPAL COMPONENTS ANALYSIS

The Statistical Package for Social Sciences (SPSS) (14) subprogram FACTOR was used for the principal components analysis. It is designed both for the factor analysis and the principal components analysis. The outputs are designed to be self-explanatory. In this example, the first 5 components account for 90% of the variance and the remaining components account for only 10% of the variance (Figure 6).

The subprogram FACTOR provides a graphical presentation (Figure 7) for the factors that have been determined by the orthogonal rotations (in this example, variance maximization rotation). In reading the graphs, one should be attentive to following three features: (1) the relative distance of a variable from the axis, (2) the direction of a variable in relation to the axis, and finally (3) clustering of variables and their relative position to each other.

03/19/80

PRINCIPAL COMPONENT ANALYSIS OF TANKS
FILE NCAPIE (CREATICA DATE = 03/19/80)

VARIABLE	EST COMMUNALITY	FACTOR	EIGENVALUE	PCT CF VAR	CLM PCT
1	1.00000	1	4.78856	47.9	47.9
2	1.00000	2	1.71316	17.1	65.0
3	1.00000	3	1.26768	12.7	77.7
4	1.00000	4	0.74450	7.4	85.1
5	1.00000	5	0.55505	5.6	90.7
6	1.00000	6	0.34472	3.4	94.1
7	1.00000	7	0.26042	2.6	96.7
8	1.00000	8	0.19890	2.0	98.7
9	1.00000	9	0.08262	0.8	99.6
10	1.00000	10	0.04438	0.4	100.0

Figure 6. Summary Table of Principal Components Analysis on Tank

PRINCIPAL COMPONENT ANALYSIS OF TANKS
 FILE NC700E (CREATION DATE = 03/15/80)

03/15/80

HORIZONTAL FACTOR 1 VERTICAL FACTOR 2

1 = 12
 2 = 14
 3 = 16
 4 = 18
 5 = 20

1 = 12
 2 = 14
 3 = 16
 4 = 18
 5 = 20

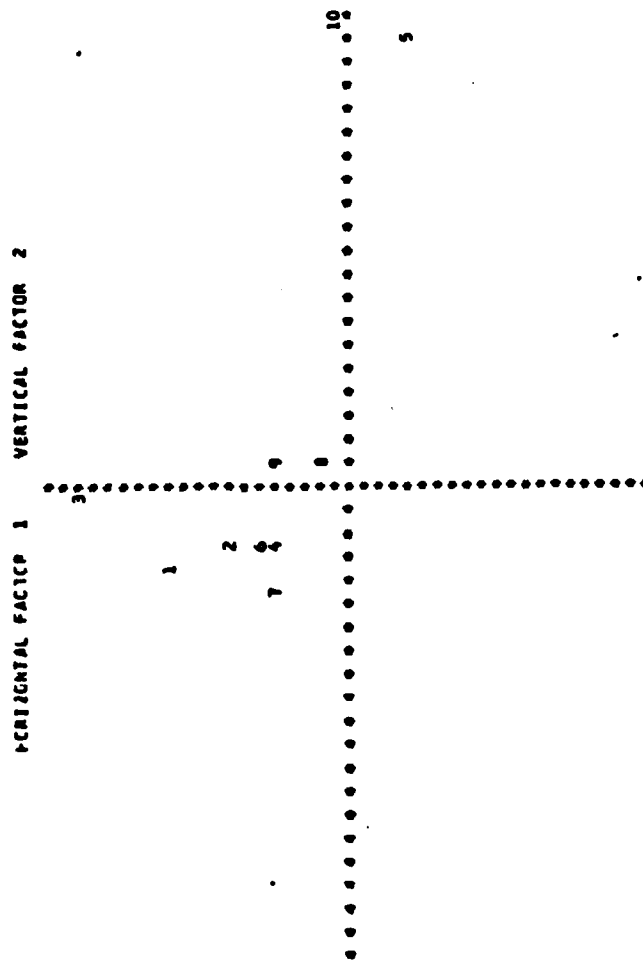


Figure 7. Graphical Presentation

03/15/80

PRINCIPAL COMPONENT ANALYSIS OF TANKS

FILE NAME: CREATION DATE = 03/15/80

FACTOR SCORE COEFFICIENTS

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8	FACTOR 9	FACTOR 10
1	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999

Figure 8. Factor Score Coefficients

For example, variables 5 (road speed) and 10 (power to engine ratio) contribute heavily to the first principal component while variables 1 (weight) and 3 (width) contribute most strongly to the second principal component. Variables 2, 4, 6, 7, 8, 9 are not as important. The weights accorded each variable in the 10 factors (principal components) are shown in Figure 8. The complete SPSS program is listed in Appendix C.

D. DISCRIMINANT ANALYSIS

The SPSS subprogram DISCRIMINANT was used to determine that function or those functions of the 10 variables that best discriminant among the four clusters determined in previous section.

The maximum number of discriminant functions to be derived is either one less than the number of groups or equal to the number of discriminating variables. This subprogram provides two measures for judging the importance of discriminant functions. One of these is the relative percentage of the eigenvalue associated with the function. It is a measure of the relative importance of the function. The sum of the eigenvalues is a measure of total variance existing in the discriminating variables. Since discriminant functions are derived in order of their importance, this process can be stopped whenever the relative percentage is judged to be too small. Of course, there is no fixed rule for deciding what is too small. In this research, we selected arbitrary, a significance level of 0.10. The output shown

in Figure 9 suggests that we therefore consider only the first two discriminant functions.

The second measure judging the importance of a discriminant function is its associated canonical correlation. The canonical correlation is a measure of association between the single discriminant function and the set of $(g-1)$ dummy variables which define the g group memberships. It tells us how closely the function and the group variable are related, which is just another measure of the function's ability to discriminate among the groups. From Figure 10, the first two discriminant functions are each highly correlated with the groups but the third has only a moderate correlation.

The next criterion for eliminating discriminant functions is to test for the statistical significance of discriminating information not already accounted for by the earlier functions. As each function is derived, starting with no (zero) functions, Wilks' lambda is computed. Lambda is an inverse measure of the discriminating power in original variables which has not yet been removed by the discriminant functions - the larger lambda is, the less is the information remaining. Lambda can be transformed into a chi-square statistic for an easy test of statistical significance. In Figure 9, Wilks' lambda was .594 after the first two functions had been derived. This corresponds to a chi-square of 8.8476 with a probability level of .1823.

TANK DISCRIMINANT

03/15/60

SUMMARY TABLE

STEP	ENTERED	ACTUAL	WORS	WILKS'	SIG.	MINIMUM	SIG.	BETWEEN GROUPS	LABEL
				LAMBDA		D SQUARE			
1	1	1	1	0.4943	0.0001	9.7233	0.0001	1	MAX ARMAMENT
2	2	2	2	0.4943	0.0000	2.7233	0.0000	2	TREACH CROSSING
3	3	3	3	0.4943	0.0000	2.7233	0.0000	3	WEIGHT CLEARANCE
4	4	4	4	0.4943	0.0000	5.6935	0.0000	4	WEIGHT CLEARANCE
5	5	5	5	0.4943	0.0000	9.8135	0.0000	5	WEIGHT CLEARANCE
6	6	6	6	0.4943	0.0000	9.8135	0.0000	6	WEIGHT CLEARANCE
7	7	7	7	0.4943	0.0000	9.8135	0.0000	7	WEIGHT CLEARANCE
8	8	8	8	0.4943	0.0000	9.8135	0.0000	8	WEIGHT CLEARANCE
9	9	9	9	0.4943	0.0000	9.8135	0.0000	9	WEIGHT CLEARANCE
10	10	10	10	0.4943	0.0000	9.8135	0.0000	10	WEIGHT CLEARANCE
11	11	11	11	0.4943	0.0000	9.8135	0.0000	11	WEIGHT CLEARANCE
12	12	12	12	0.4943	0.0000	9.8135	0.0000	12	WEIGHT CLEARANCE

CLASSIFICATION FUNCTION COEFFICIENTS

PATIENTS	1	2	3	4
1	-1.1250	-1.1250	-1.1250	-1.1250
2	-1.1250	-1.1250	-1.1250	-1.1250
3	-1.1250	-1.1250	-1.1250	-1.1250
4	-1.1250	-1.1250	-1.1250	-1.1250
5	-1.1250	-1.1250	-1.1250	-1.1250
6	-1.1250	-1.1250	-1.1250	-1.1250
7	-1.1250	-1.1250	-1.1250	-1.1250
8	-1.1250	-1.1250	-1.1250	-1.1250
9	-1.1250	-1.1250	-1.1250	-1.1250
10	-1.1250	-1.1250	-1.1250	-1.1250
11	-1.1250	-1.1250	-1.1250	-1.1250
12	-1.1250	-1.1250	-1.1250	-1.1250

CANONICAL DISCRIMINANT FUNCTIONS

FUNCTION	EIGENVALUE	PERCENT OF VARIANCE	CUMULATIVE PERCENT	CANONICAL CORRELATION	AFTER FUNCTION	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
1	1.20597	42.38	42.38	0.944927	1	0.4943	72.507	14	0.0000
2	1.20597	42.38	84.76	0.882162	2	0.114253	35.079	14	0.0000
3	1.20597	42.38	100.00	0.856909	3	0.114253	8.276	6	0.1023

0 MARKS THE 3 CANONICAL DISCRIMINANT FUNCTION(S) TO BE USED IN THE REPAIRING ANALYSIS.

Figure 9. Summary Table of Discriminant Analysis on Tank

TNR DISCRIMINANT

STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

	FUNC 1	FUNC 2	FUNC 3
PC	0.55115	-0.26099	-0.07560
PC	-0.27182	-0.14927	-0.54022
PC	-0.04922	-0.17521	-0.08222
PC	-0.11118	-0.11611	-0.02267
PC	-0.04922	-0.11118	-0.02267
PC	-0.04922	-0.11118	-0.02267
PC	-0.04922	-0.11118	-0.02267
PC	-0.04922	-0.11118	-0.02267
PC	-0.04922	-0.11118	-0.02267
PC	-0.04922	-0.11118	-0.02267

RELATIONSHIPS BETWEEN CANONICAL DISCRIMINANT FUNCTIONS ARE DISCRIMINATING VARIABLES ARE CERVED BY THE FUNCTION WITH LARGEST CORRELATION AND THE MAGNITUDE OF THAT CORRELATION.

	FUNC 1	FUNC 2	FUNC 3
PC	0.26099	-0.14927	0.01268
PC	-0.27182	-0.14927	0.01268
PC	-0.04922	-0.17521	0.23805
PC	-0.11118	-0.11611	-0.14927
PC	-0.04922	-0.11118	-0.14927
PC	-0.04922	-0.11118	-0.14927
PC	-0.04922	-0.11118	-0.14927
PC	-0.04922	-0.11118	-0.14927
PC	-0.04922	-0.11118	-0.14927
PC	-0.04922	-0.11118	-0.14927

LASTARBEFCEZEE CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

	FUNC 1	FUNC 2	FUNC 3
PC	0.1245922	0.37626720-C1	-0.9502430-02
PC	-0.1245922	-0.37626720-C1	0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232
PC	-0.1245922	-0.37626720-C1	-0.276232

Figure 10. Canonical Discriminant Function Coefficients

This means that a lambda of this magnitude or smaller has a .1823 probability of occurring due to the chances of sampling even if there was no further information to be accounted for by a third function in the population. Clearly, a third function is not statistically significant in this case.

The standardized discriminant function coefficients corresponding to the values of the v_{ij} 's discussed in the previous section are used to compute the discriminant score for a case (observation) in which the original discriminating variables are in standard form. The discriminant score is computed by multiplying each discriminating variable by its corresponding coefficient and adding together these products. There is a separate score for each observation on each function. The coefficients have been derived in such a way that the discriminant scores produced are in standard form.

When the sign is ignored, each standard discriminant function coefficient represents the relative contribution of its associated variable to that function. The sign merely denotes whether the variable is making a positive or negative contribution.

A graphical presentation is shown in Figure 11 using the first and the second canonical discriminant function as the axis. From this scatterplot, we can easily see that Soviet tanks (labelled 1) are well distinguished from the all of the others using only the first two discriminant

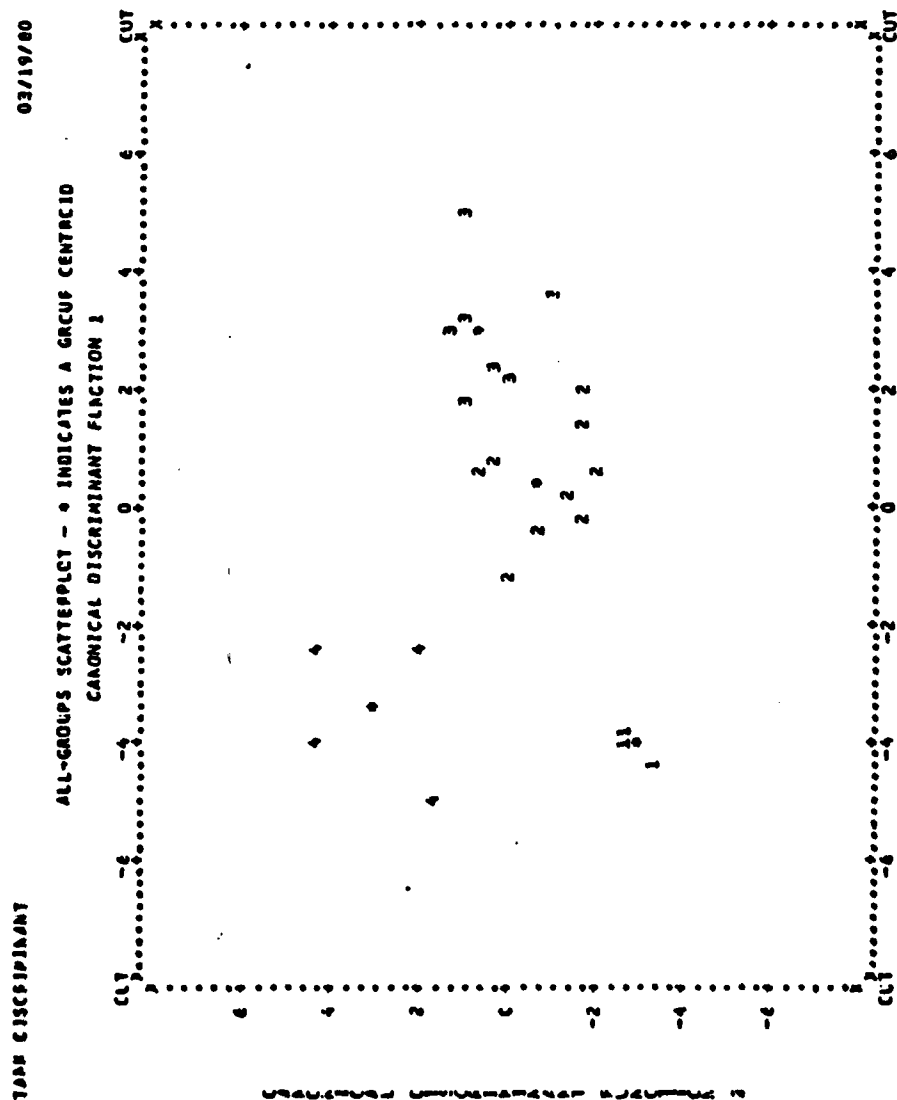


Figure 11. Scatter Plot for all Groups

functions. Also, all the lightweight tanks are clearly separated from the others. The distinction between groups 2 and 3 is also clear though not separated from each other as much as from groups 1 and 4. The complete SPSS program for the discriminant analysis is listed in Appendix D.

VII. CONCLUSION

The multivariate analysis techniques of cluster analysis, principal components analysis and discriminant analysis are useful in real world problems for examining observations on each of several dimension. Each of the techniques is related mathematically to the others, and each complements the other in explaining the data.

Computer software is readily available in many sources. The software used in this thesis for hierarchical clustering, principal components analysis, and discriminant analysis was from the IMSL package and SPSS. For nonhierarchical clustering, we used the FORTRAN program developed by McRae (16). All of this software is readily available and documented at the Naval Postgraduate School.

Appendix A:

HIERARCHICAL CLUSTERING

THIS IS A PROGRAM FOR HIERARCHICAL CLUSTERING

```

VARIABLE DESCRIPTION
ND INPUT NUMBER OF DATA POINTS TO BE CLUSTERED.
ND-1 CLUSTERED ARE FORMED CONSECUTIVELY
FROM ND+1 TO NC+(NC-1). ND MUST BE GREATER THAN 2.
ND IN THE RANGE 2,3,....,500 IS ALLCWD
(SEE REMARKS)
INPUT OPTIONS VECTOR OF LENGTH 2
IOPT(1)=0 IMPLIES SINGLE-LINKAGE IS DESIRED.
OTHERWISE COMPLETE-LINKAGE IS DESIRED.
IOPT(2)=0 IMPLIES THE SIMILARITIES ARE DISTANCE-
LIKE (I.E. SMALLER IMPLIES CLOSER). OTHERWISE
THE SIMILARITIES ARE ASSUMED TO BE POSITIVE
CORRELATION-LIKE SIMILARITIES. (SEE REMARKS)
INPUT/OUTPUT VECTOR OF LENGTH (NC+1)*NC/2
CONTAINING THE SIMILARITY MATRIX IN SYMMETRIC
STORAGE MODE. FOR I GREATER THAN J,
XSIM((I-1)*I+J) CONTAINS THE SIMILARITY OF
THE I-TH AND J-TH DATA POINTS. ON INPUT, THE
DIAGONAL ELEMENTS (SIMILARITY OF ITSELF) ARE
ARBITRARY AND DO NOT NEED TO BE DEFINED.
XSIM IS DESTROYED ON OUTPUT.
CLEVEL OUTPUT VECTOR OF LENGTH ND-1. CLEVEL(K) CONTAINS THE
SIMILARITY LEVEL AT WHICH CLUSTER NC+K WAS FORMED.
DIMENSION XSIM(500), ICPT(2), CLEVEL(500), TITLE(20),
ICRSCN(500), IPTR(100), ICUT(100), INCL(4), CLVSK(100), NCLST(100),
ZLEFT(100), STARST(100), YSIM(2)
THIS WOULD BE THE INPUT VECTOR FOR SUPERPROGRAM USTREE.
ICLSCN OUTPUT VECTOR OF LENGTH NC-1. CLUSTER
NUMBER NC+K WAS FORMED BY MERGING CLUSTER ICLSCN(K)
WITH CLUSTER ICRSCN(K). VECTOR FOR SUPERPROGRAM USTREE.
THIS WOULD BE THE INPUT VECTOR FOR SUBPROGRAM USTREE.
OUTPUT VECTOR OF RIGHTSONS OF LENGTH NC-1. THE
RIGHTSON OF CLUSTER NC+K IS CONTAINED IN ICRSCN(K).
THIS WOULD BE THE INPUT VECTOR FOR SUPERPROGRAM USTREE.
WORK VECTOR OF LENGTH ND.
IPTR ERROR PARAMETER. (CUTPUT)
IER TERMINAL ERROR
IER=129 INDICATES ND WAS LESS THAN 3.
INPUT VECTOR OF LENGTH 4. IND(1) CONTAINS WHEN I=1,
THE HEAD NODE OF THE SUBTREE TO BE PRINTED. (REMARKS)
MUST BE BETWEEN NC AND 2*NC. EXCLUSIVELY. (REMARKS)
I=2, NUMBER BETWEEN PRINTABLE SPACES PER LINE ON THE
OUTPUT (PRINTER) DEVICE. IND(2) MUST EXCEED 4.

```

```

I=3, NUMBER OF HORIZONTAL SLICES OF TREE DESIRED
TO PROVIDE THE NECESSARY DETAIL.
INC(3) MUST BE POSITIVE.
I=4, NUMBER OF FILLER LINES PRINTED BETWEEN NCDE
LINES (1 USUALLY SUFFICIENT).
INPUT VECTOR OF LENGTH 2 CONTAINING THE INTERVAL CN
THE VECTOR SCALE USED TO CUT THE TREE
(SEE VECTOR CLEVEL). LEVEL YSIM(1) IS WHERE THE
TERMINAL NODES (FOR IMSCUSTERING ROUTINES, THE
DATA POINTS) ARE PRINTED. THE OTHER INTERVAL
ENDPOINT YSIM(2) SHOULD INCLUDE THE LEVEL FOR THE
HEAD NCDE (INC(1)). SEE REMARKS.

NAMELIST /NAME1/XSIM VECTOR OF LENGTH INC(2)-4.
ICUT WORK VECTOR OF LENGTH ND.
CLVLSK WORK VECTOR OF LENGTH ND.
NCLRST WORK VECTOR OF LENGTH ND.
LEFTRT WORK VECTOR OF LENGTH ND.
STARST WORK VECTOR OF LENGTH ND.

YSIM

```

REMARKS 1. THE DATA CLUSTERS ARE NUMBERED 1 TO NC. THE ND-1 CLUSTERS FORMED BY MERGING ARE NUMBERED ND+1 TO NC+(ND-1) AND DECREASE IN SIMILARITY, MAKING IT EASY TO IDENTIFY THE MOST SIMILAR CLUSTERS.

2. SIMILARITIES GENERALLY SHOULD BE NONNEGATIVE. RAW CORRELATIONS TAKE CN VALUES R WHERE R LIES IN THE CLOSED INTERVAL (-1, 1) AND SHOULD BE MADE POSITIVE. IF R=1 AND R=-1 BOTH MEAN HIGH SIMILARITY, THEN THE TRANSFORMATION, $RR = 1 + R$ EQUAL TO THE ABSOLUTE VALUE OF R, IS EQUIVALENT TO R SQUARED, ARE APPROPRIATE. IF R=-1 REPRESENTS A VERY LOW SIMILARITY, THEN THE TRANSFORMATION, $RR = 1 - R$, BECOMES A DISTANCE-LIKE SIMILARITY.

3. NOTE THAT THE ORIGINAL DATA MATRIX OF THE USER DOES NOT ENTER CLINK, ALLOWING THE USER TO DEFINE, VIA XSIM, WHATEVER MEASURE OF SIMILARITY SEEMS MOST APPROPRIATE.

4. THE TREE MAY BE TOO LARGE TO FIT IN INC(2) SPACES REPRESENTING THE INTERVAL (YSIM(1), YSIM(2)). IF SCALSTREE CAN PRINT THE TREE IN INC(3) SLICES OF THIS INTERVAL WHICH MAY BE CUT AND TAPED TOGETHER.

5. TO PRINT THE ENTIRE TREE FROM IMSL SUBROUTINE CLINK, THE HEAD NODE INC(1) = NC+(ND-1).

6. OUTPUT IS WRITTEN TO THE UNIT SPECIFIED BY IMSL ROUTINE UGETIO. SEE THE UGETIO DOCUMENT FOR DETAILS.

25 JAN 1973

SIMILARITY ARE NEGATIVE CORRELATION-LIKE.

ALPHABETIC PRESERVATIONS IN DISM ARE 351

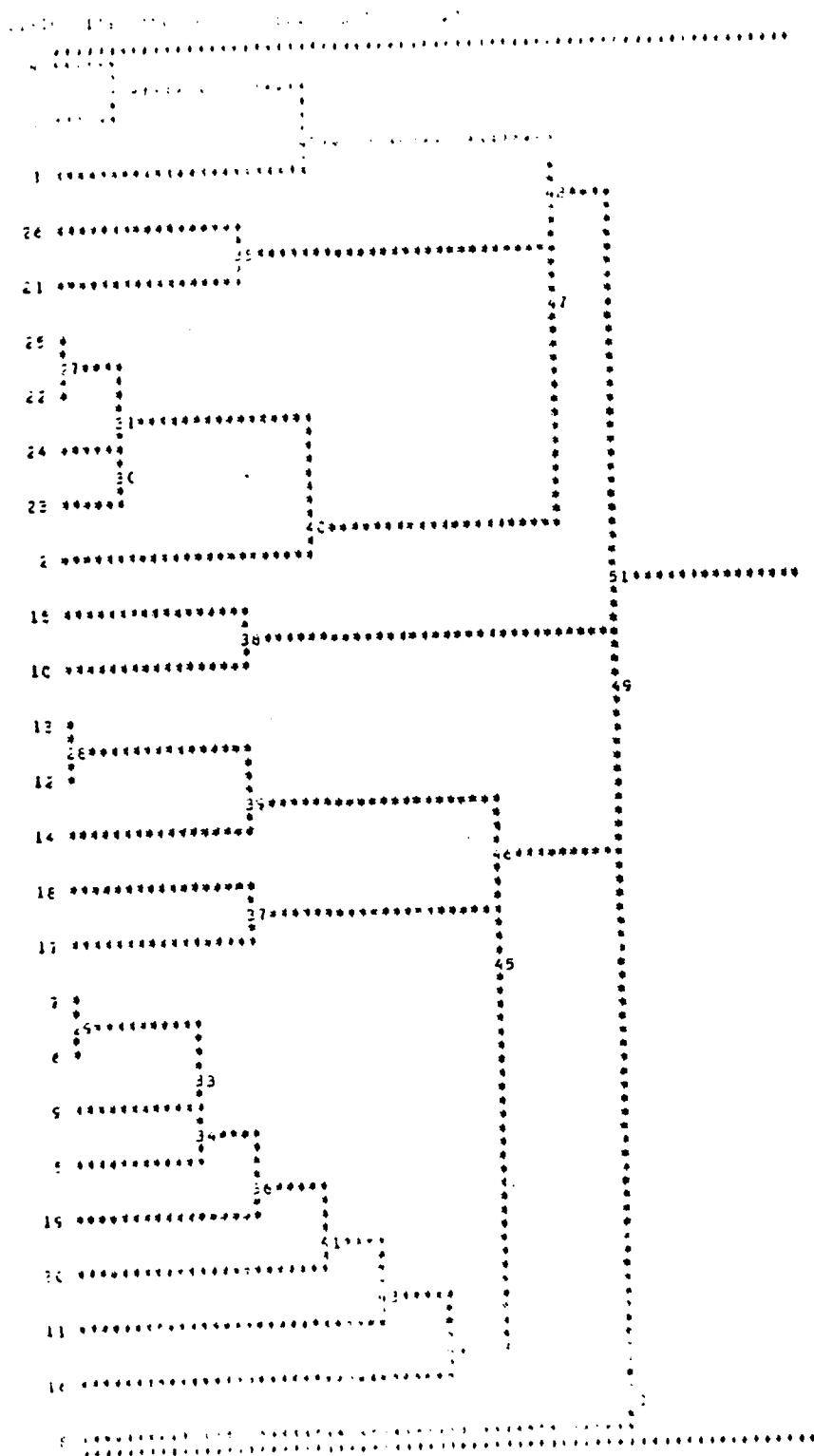
INITIAL DISTANCE MATRIX

[illegible]

ICRSON
22
126
230
335
55
219
17
10
14
20
11
16
44
45
40
47
46
8
50

ICLSCN
172244532918531624177542848

[illegible]



Appendix B:

K-MEANS ITERATIVE CLUSTERING PROGRAM BY D.J. MCRAE
 THIS PROGRAM ENABLES THE USER TO CLUSTER A DATA MATRIX
 UP TO SIZE (600 X 20) INTO A MAXIMUM OF 20 CLUSTERS.
 SEVERAL OPTIONS ARE AVAILABLE WITH RESPECT TO THE
 METHOD OF COMPUTATION AND CRITERIA TO ACCOMPLISH THE
 CLUSTERING. THE PROGRAM MAY BE USED ON EITHER
 THE CP/CMS TERMINALS OR BY THE CARD READER. THE APPROPRIATE
 METHOD FOR EACH IS DESCRIBED BELOW.

CP/CMS TERMINAL USE
 THE PROGRAM SHOULD BE READ INTO CP AS IS, BUT
 PRECEDED BY THE FOLLOWING CARDS, STARTING IN COLUMN ONE

CF67USERID XXXXG
 CFFLINE READ RRREAD FCRTAN

THE DATA MATRIX CAN BE PLACED IN THE COMPILE
 EITHER BY USE OF THE CFFLINE READ CR BY TYPING IN THE
 INFORMATION ON THE TERMINAL. IF CFFLINE READ IS USED THE DATA
 MUST BE IN THE SPECIFIED FORMAT PRECEDED BY THE FOLLOWING
 TVC CARDS, STARTING IN COLUMN ONE:

CF67USERID XXXXG
 CFFLINE READ FILE FT04FYYY

THE PROGRAM CAN THEN BE EXECUTED BY FIRST ISSUING THE

CMS COMMAND
 FOLLOWED BY

F RRREAD
 \$ RRREAD

CS/CARD READER USE

THE PROGRAM MUST CONTAIN THE FOLLOWING INFORMATION
 IN THE FOLLOWING ORDER:

STANDARD GREEN JOB CARD, TIME=(AS DESIRED)
 //EXEC FORTCLG,REGION=60=20CK
 //FCRT,SYSD DD *
 //MAIN PROGRAM
 /*
 //GC,SYSD DD *
 //DATA DECK
 /*

RRRCC12C
 RRR0001C
 RRRCC020
 RRRCC03C
 RRR00040
 RRR0005C
 RRR00060
 RRRCC07C
 RRR00080
 RRRCC09C
 RRRCC10C
 RRRCC11C
 RRRCC13C
 RRRCC140
 RRRCC15C
 RRR00160
 RRRCC17C
 RRR00180
 RRRCC19C
 RRR00200
 RRRCC21C
 RRR00220
 PRR00240
 RRR00270
 RRR00280
 RRR00330
 RRR0034C
 RRR0036C
 RRRCC37C
 RRRCC4CC
 RRR0041C
 RRRCC42C
 RRR00430
 RRR00440
 RRR0045C
 RRR0046C
 RRR0047C
 RRR0048C

RRR000490
RRR000500
RRR000510
RRR000520
RRR000530
RRR000540
RRR000550
RRR000560
RRR000570
RRR000580
RRR000590
RRR000600
RRR000610
RRR000620
RRR000630
RRR000640
RRR000650
RRR000660
RRR000670
RRR000680
RRR000690
RRR000700
RRR000710
RRR000720
RRR000730
RRR000740
RRR000750
RRR000760
RRR000770
RRR000780
RRR000790
RRR000800
RRR000810
RRR000820
RRR000830
RRR000840
RRR000850
RRR000860
RRR000870
RRR000880
RRR000890
RRR000900
RRR000910
RRR000920
RRR000930
RRR000940
RRR000950
RRR000960

DATA INPUT DECK

THE FIRST CARD OF THE DATA DECK IS THE TITLE CARD.
IT MAY CONTAIN ANY ALPHA NUMERIC TITLE IN COLUMNS 1 - 80.

THE SECOND CARD IS THE PROBLEM CARD. IT CONTAINS
INTEGERS IN THE FIRST 13 COLUMNS IN THE FOLLOWING MANNER:
CCL 1-4: NUMBER OF OBSERVATIONS
CCL 5-6: NUMBER OF VARIABLES PER OBSERVATION
CCL 7-8: NUMBER OF CLUSTERS DESIRED
CCL 9: CRITERION TO BE USED IN THE EVALUATION:
1 = TRACE W
2 = DETERMINANT W
3 = LARGEST EIGENVALUE OF W(INVERSE)*B
4 = TRACE W(INVERSE)*B
CCL 10: STANDARDIZATION PARAMETER:
C = THE PROGRAM WILL NOT STANDARDIZE DATA
1 = THE PROGRAM WILL STANDARDIZE DATA
CCL 11: DISTANCE PARAMETER:
0 = EUCLIDIAN DISTANCE
1 = SCALED EUCLIDIAN DISTANCE
2 = MAHALANOBIS DISTANCE
CCL 12: DATA PARAMETER:
C = DATA IS
1 = NO DATA
CCL 13: TIMING PARAMETER:
0 = NO OBSERVATIONS CONSIDERED IN
1 - 8 = AN INCREASING NUMBER OF OBSERVATIONS
9 = ALL OBSERVATIONS USED
NOTE: THE TIMING PARAMETER DETERMINES HOW EXTENSIVE THE
DOUBLE CHECK OF THE CLUSTER SOLUTION IS TO BE.
NORMALLY '9' IS USED UNLESS THE DATA MATRIX IS LARGE

THIS CARD MUST BE IN (14,212,511) FORMAT

THE THIRD CARD IS THE VARIABLE FORMAT CARD. THIS CARD
MUST BE PRESENT UNLESS THE DATA PARAMETER IN THE PROBLEM
CARD IS '1'. THE FORMAT STATEMENT IS THAT TYPE WHICH THE DATA
CARDS ARE IN. FOR EXAMPLE IF THE DATA CARD HAS DATA:
ABCD 1.2 1.55
STARTING IN COL 1, THE VARIABLE FORMAT CARD WOULD
CONTAIN THE FOLLOWING STARTING IN CCL 1:
(1A4,4X,F3.1,4X,F4.2)

THE NEXT CARDS ARE THE DATA CARDS. THEY MAY CONTAIN
DATA IN COLUMNS 1 - 72.

CC

```

CCCCCCCCCCCC
    FCR REANALYSIS OF THE SAME DATA, ACC A NEW TITLE CARD
    FOLLOWED BY A NEW PROBLEM CARD, WITH A '1' IN COLUMN 11
    CF THE NEW PROBLEM CARD.
    FCR ANALYSIS OF NEW DATA, SIMPLY PLACE THE NEW DATA
    SET, INCLUDING NEW TITLE CARD, PROBLEM CARD, AND FORMAT CARD,
    AFTER THE PREVIOUS SET
    FCR THE PROGRAM TO EXIT NORMALLY, TWO BLANK CARDS MUST FOLLOW
    THE LAST DATA BATCH.
    CCMCN NOBS, NVAR, NGPS, ICRIT, NOSTAN, IDIST, IFINE, KTIME, ICENT(600),
    IDATA(600), T(20,20), B(20,20), W(20,20), W(20,20), SVCEN(20,20),
    2 IDATA(600), NISV(20), VMEAN(20), SD(20), XVEC(20), YVEC(20),
    3 NISVT(20), SVCENT(20,20), IDATA(600), VEC(20,20), EIG(20)
CCCCCCCCCCCC
    DESCRIPTION OF COMMON AREA:
    NOBS = NUMBER OF OBSERVATIONS
    NVAR = NUMBER OF VARIABLES
    NGPS = NUMBER OF CLUSTERS
    ICRIT = CRITERION TO BE OPTIMIZED (SEE ABOVE)
    NOSTAN = DISTANCE PARAMETER (SEE ABOVE)
    IDIST = ESCAPE PARAMETER: IF IFINE IS SET EQUAL TO '1',
    IFINE = SOMETHING IS WRONG AND THE APPROPRIATE ERROR
    MESSAGE IS PRINTED OUT; THE PROGRAM GOES ON TO THE
    NEXT PROBLEM
    KTIME = TIMING PARAMETER (SEE ABOVE)
    IDENT = OBSERVATION IDENTIFICATIONS (A4 FORMAT): READ FROM
    EACH DATA CARD THE DATA MATRIX IS STORED
    DATA = AREA IN WHICH THE DATA MATRIX IS STORED
    T = CROSS-CLUSTERS MATRIX
    B = BETWEEN-CLUSTERS MATRIX
    W = WITHIN-CLUSTERS MATRIX
    W(20,20) = CHCLESKY FACTOR OF THE WITHIN-CLUSTERS MATRIX
    SVCEN = CLUSTER CENTERS (MEANS)
    IDATA = CLUSTER IDENTIFICATION FOR EACH OBSERVATION
    NISV = CLUSTER SIZES (NUMBER OF OBSERVATIONS IN EACH CLUSTER)
    VMEAN = VARIABLE MEANS
    SD = VARIABLE STANDARD DEVIATIONS
    XVEC = TEMPORARY STORAGE
    YVEC = TEMPORARY STORAGE
    NAN=24
    BT = TEMPORARY STORAGE FOR BETWEEN-CLUSTERS MATRIX
    NISVT, SVCENT, IDATA: TEMPORARY STORAGE SERVING THE
    SAME FUNCTIONS AS NISV, SVCEN, AND IDATA
CCCCCCCCCCCC

```

```

RRRC057C
RRRC058C
RRRC059C
RRRC060C
RRRC061C
RRRC062C
RRRC063C
RRRC064C
RRRC065C
RRRC066C
RRRC067C
RRRC068C
RRRC069C
RRRC070C
RRRC071C
RRRC072C
RRRC073C
RRRC074C
RRRC075C
RRRC076C
RRRC077C
RRRC078C
RRRC079C
RRRC080C
RRRC081C
RRRC082C
RRRC083C
RRRC084C
RRRC085C
RRRC086C
RRRC087C
RRRC088C
RRRC089C
RRRC090C
RRRC091C
RRRC092C
RRRC093C
RRRC094C
RRRC095C
RRRC096C
RRRC097C
RRRC098C
RRRC099C
RRRC100C
RRRC101C
RRRC102C
RRRC103C
RRRC104C
RRRC105C
RRRC106C
RRRC107C
RRRC108C
RRRC109C
RRRC110C
RRRC111C
RRRC112C
RRRC113C
RRRC114C
RRRC115C
RRRC116C
RRRC117C
RRRC118C
RRRC119C
RRRC120C
RRRC121C
RRRC122C
RRRC123C
RRRC124C
RRRC125C
RRRC126C
RRRC127C
RRRC128C
RRRC129C
RRRC130C
RRRC131C
RRRC132C
RRRC133C
RRRC134C
RRRC135C
RRRC136C
RRRC137C
RRRC138C
RRRC139C
RRRC140C
RRRC141C
RRRC142C
RRRC143C

```

```

C      VEC = EIGENVECTORS
C      EIG = EIGENVALUES
C
C      1CC CALL PRELIM (CRIT)
C      IF (IFINE.EC.1) GC TO 400
C      IF (IFINE.EC.2) GC TO 500
C      CALL IRANDST (CRIT)
C      IF (IFINE.EC.1) GC TO 400
C      CALL KMEANS (CRIT)
C      IF (IFINE.EC.1) GC TO 400
C      CALL ISWICH (CRIT)
C      GC TO 100
C      IFINE GOT SET TO 1: IF DATA HAS BEEN READ IN, RESET DATA AND
C      GC TC NEXT PROBLEM
C      4CC GC 410 I=1, NBOBS
C      GC 410 J=1, NBOBS
C      IF (NGSTAN.EQ.1) DATA(I,J) = DATA(I,J) * SD(J)
C      DATA(I,J) = DATA(I,J) + VMEAN(J)
C      5CC GC TO 100
C      ALL DONE: WRITE CLT LAST MESSAGE AND EXIT
C      5CC WRITE (6,50C)
C      STOP
C      ENCL
C      5CC FORMAT (10 END OF ANALYSES.)
C
C      SUBROUTINE PRELIM (CRIT)
C
C      THIS SUBROUTINE MAKES THE PRELIMINARY CALCULATIONS.
C      IT INPUTS THE DATA, CALCULATES THE MEANS AND VARIANCES FOR EACH
C      VARIABLE, STANDARDIZES (CONVERTS TO Z-SCORES) EACH VARIABLE IF
C      REQUESTED, AND CALCULATES THE CROSS-PRODUCTS MATRIX
C
C      COMMON NOBS,NVARS,NGPS,ICRIT,NOSTAN,IDIST,IFINE,KTIME,IDENT(6C0),
C      1 DATAT(6C0,20),T(20,20),B(20,20),WFC1(2C,20),SVCE1(2C,2C),
C      2 ICATA(6C0),NISV(2C),VMEAN(20),SC(2C),XVEC(2C),YVEC(20),ET(20,20),
C      3 NISVT(20),SVCENT(20,20),IDATAT(6C0),VEC(20,20),EIG(2C)
C      DIMENSION TITLE(20),IFMT(20),VAR(2C)
C
C      INPUT SECTION: READ IN TITLE CARD, PROBLEM CARD, OPTIONAL CARDS,
C      FCRMAT CARD, AND DATA CARDS
C      WRITE OUT SOLUTION SPECIFICATIONS
C
C      IFINE = 0
C      READ (5,900) (TITLE(I),I=1,20)
C      WRITE (6,503) (TITLE(I),I=1,20)
C      READ (5,501) NOBS,NVARS,NGPS,ICRIT,NOSTAN,ICIST,NCLDATA,KTIME
C      IF (NOBS.EQ.0) GC TC 820

```

```

RRR0144C
RRR0146C
RRR0147C
RRRC148C
RRR0149C
RRR0150C
RRR0151C
RRR0152C
RRR0153C
RRR0154C
RRRC155C
RRR0156C
RRRC157C
RRR0158C
RRRC159C
RRR0160C
RRRC161C
RRR0163C
RRR0164C
RRR0165C
RRR0166C
RRRC167C
RRR0168C
RRRC169C
RRRC170C
RRR0171C
RRR0172C
RRR0173C
RRRC174C
RRR0175C
RRR0176C
RRR0177C
RRR0178C
RRR0179C
RRR0180C
RRRC181C
RRR0182C
RRR0183C
RRRC184C
RRR0185C
RRRC186C
RRRC187C
RRR0188C
RRR0189C
RRR0190C
RRR0191C
RRR0192C
RRR0193C
RRR0194C
RRR0195C
RRR0196C
RRR0197C
RRR0198C
RRR0199C
RRR0200C

```



```

RRRC238C
RRRC239C
RRRC240C
RRRC241C
RRRC242C
RRRC243C
RRRC244C
RRRC245C
RRRC246C
RRRC247C
RRRC248C
RRRC249C
RRRC250C
RRRC251C
RRRC252C
RRRC253C
RRRC254C
RRRC255C
RRRC256C
RRRC257C
RRRC258C
RRRC259C
RRRC260C
RRRC261C
RRRC262C
RRRC263C
RRRC264C
RRRC265C
RRRC266C
RRRC267C
RRRC268C
RRRC269C
RRRC270C
RRRC271C
RRRC272C
RRRC273C
RRRC274C
RRRC275C
RRRC276C
RRRC277C
RRRC278C
RRRC279C
RRRC280C
RRRC281C
RRRC282C

```

```

CC 105 J=1,NVARS
VMEAN(J)=0.0
105 VAF(J)=0.0
CC 110 I=1,NBS
CC 111 J=1,NVARS
110 VMEAN(J)=VMEAN(J)+DATA(I,J)/NBS
CC 120 I=1,NBS
CC 121 J=1,NVARS
DATA(I,J)=DATA(I,J)-VMEAN(J)
120 VAR(J)=VAF(J)+(DATA(I,J)**2)/(NBS-1)
CC 125 J=1,NVARS
IF (VAR(J).LE.0.00001) MFLAG=1
125 SC(J)=SORT (VAR(J))
IF (MFLAG.EQ.1) GO TO 825

CC CALCULATE T = X'X, THE CRCSS-PRODUCTS MATRIX
CC
131 CC 135 K=1,NVARS
CC 136 J=K,NVARS
135 T(K,J)=0.0
CC 140 K=1,NVARS
CC 141 J=K,NVARS
140 T(K,J)=T(K,J)+DATA(I,K)*DATA(I,J)

CC OUTPUT SECTION: WRITE OUT MEANS, STANDARD DEVIATIONS, AND CRCSS-
CC PRODUCTS MATRIX
CC
WRITE (6,90E) (VMEAN(J),J=1,NVARS)
WRITE (6,90E)
WRITE (6,90E) (SC(J),J=1,NVARS)
WRITE (6,91E)
CC 810 I=1,NVARS
WRITE (6,99E) (T(J,I),J=1,I)
CC CONTINUE
810 IF (NCSTAN.EQ.0) GO TO 180

CC STANDARDIZE IF REQUESTED
CC
CC 160 J=1,NVARS
CC 161 I=1,NBS
DATA(I,J)=DATA(I,J)/SD(J)
150 CC 160 K=J,NVARS
CC 161 K=J,NVARS
T(J,K)=(1.0/SD(J))*T(J,K)*(1.0/SD(K))
160 WRITE (6,70E)
CC 700 I=1,NBS
WRITE (6,70E) (DATA(I,L),L=1,NVARS)

```

R R R C E I 190
R R R R O E I 170
R R R R O E I 160
R R R R O E I 150
R R R R F C E I 140
R R R R O E I 130
R R R R O E I 120
R R R R O E I 110
R R R R O E I 100
R R R R O E I 090
R R R R C E I 80
R R R R C E I 70
C 50

RRRRC320C
 RRRRC321C
 RRRRC322C
 RRRRC323C
 RRRRC324C
 RRRRC325C
 RRRRC326C
 RRRRC327C
 RRRRC328C
 RRRRC329C
 RRRRC330C
 RRRRC331C
 RRRRC332C
 RRRRC333C
 RRRRC334C
 RRRRC335C
 RRRRC336C
 RRRRC337C
 RRRRC338C
 RRRRC339C
 RRRRC340C
 RRRRC341C
 RRRRC342C
 RRRRC343C
 RRRRC344C
 RRRRC345C
 RRRRC346C
 RRRRC347C
 RRRRC348C
 RRRRC349C
 RRRRC350C
 RRRRC351C
 RRRRC352C
 RRRRC353C
 RRRRC354C
 RRRRC355C
 RRRRC356C
 RRRRC357C
 RRRRC358C
 RRRRC359C
 RRRRC360C
 RRRRC361C
 RRRRC362C
 RRRRC363C
 RRRRC364C
 RRRRC365C
 RRRRC366C
 RRRRC367C

AD-A086 521

NAVAL POSTGRADUATE SCHOOL MONTEREY CA
CLASSIFICATION TECHNIQUES FOR MULTIVARIATE DATA ANALYSIS. (U)
MAR 80 J K LEE

F/6 12/1

UNCLASSIFIED

NL

2 of 2
ALL
ACQUIS.



END
DATE
FILMED
8-80
DTIC

```

RRRC368G
RRRC3690
RRRC3700
RRRC3710
RRRC3720
RRRC3730
RRRC3740
RRRC3750
RRRC3760
RRRC377C
RRRC3780
RRRC3790
RRRC3800
RRRC3810
RRRC3820
RRRC3830
RRRC3840
RRRC3850
RRRC3860
RRRC3870
RRRC3880
RRRC3890
RRRC3900
RRRC3910
RRRC3920
RRRC3930
RRRC3940
RRRC3950
RRRC3960
RRRC3970
RRRC3980
RRRC3990
RRRC4000
RRRC4010
RRRC4020
RRRC4030
RRRC4040
RRRC4050
RRRC4060
RRRC4070
RRRC4080
RRRC4090
RRRC4100
RRRC4110
RRRC4120
RRRC4130
RRRC4140
RRRC4150

```

```

215 ICATAT(I) = M
C CONTINUE
C ASSIGN OBSERVATION TO CLOSEST CLUSTER AND UPDATE THAT CLUSTER
C CENTER
C K = IDATAT(I)
NISVT(K) = NISVT(K) + 1
CC 225 J = 1, NVARS
225 SVCENT(K,J) = ((NISVT(K)-1)*SVCENT(K,J)+DATA(I,J))/NISVT(K)
C ILIST = ITEPP
C ENCLCF DC FOR EACH OBSERVATION
C RECALCULATE CLUSTER CENTERS TO ELIMINATE INITIAL RANCC# CBSERVA-
C TICS
CC 230 M = 1, NGPS
NISVT(M) = NISVT(M) - 1
CC 230 J = 1, NVARS
230 SVCENT(M,J) = 0.0
C M = IDATAT(I)
CC 235 J = 1, NVARS
235 SVCENT(M,J) = SVCENT(M,J) + DATA(I,J)/NISVT(M)
C CALCULATE THE B AND W MATRICES
C CALCULATE THE CRITERION VALUE
C ICIR=2
IF (ICRIT.EC.1) ICIR=1
CALL WCALC (SVCENT,NISVT,NGPST,IOIR)
IF (IFINE.EC.1) RETURN
CALL CRITON (CRIT)
IF (IFINE.EC.1) RETURN
C WHICH INITIAL CONFIGURATION
IF (IBSRT.GT.1) GO TO 250
C FIRST INITIAL CONFIGURATION: BCRIT IS THE BEST CRITERION
245 BCRIT = CRIT
SAVE CLUSTER SIZES (NISV), CLUSTER CENTERS (SVCENT), CLUSTER
LISTS (LSTSV), AND OBSERVATION IDs (IDATA)
CC 247 M = 1, NGPS
NISVT(M) = NISVT(M)
CC 247 L = 1, NVARS
247 SVCENT(M,L) = SVCENT(M,L)
CC 249 I = 1, NOBS
249 ICATAT(I) = IDATAT(I)
CC TO 260
C SECOND OR THIRD INITIAL CONFIGURATION
250 IF (CRIT.GE.BCRIT) GO TO 260
255 BCRIT = CRIT
CC 257 M = 1, NGPS
NISVT(M) = NISVT(M)

```

RRRRC4160
RRRRC417C
RRRRC4180
RRRRC419C
RRRRC4200
RRRRC4210
RRRRC4220
RRRRC4230
RRRRC4240
RRRRC4250
RRRRC426C
RRRRC4270
RRRRC4280
RRRRC4290
RRRRC4300
RRRRC4310
RRRRC4320
RRRRC433C
RRRRC4340
RRRRC4350
RRRRC4360
RRRRC4370
RRRRC4380
RRRRC439C
RRRRC4400
RRRRC4410
RRRRC4420
RRRRC4430
RRRRC444C
RRRRC4450
RRRRC4460
RRRRC4470
RRRRC4480
RRRRC4490
RRRRC4500
RRRRC4510
RRRRC4520
RRRRC4530
RRRRC4540
RRRRC4550
RRRRC4560
RRRRC4570
RRRRC4580
RRRRC4590
RRRRC4600
RRRRC4610
RRRRC462C
RRRRC4630

```

RRRC4640
RRRC4650
RRRC4660
RRRC4670
RRRC4680
RRRC4690
RRRC4700
RRRC4710
RRRC4720
RRRC4730
RRRC4740
RRRC4750
RRRC4760
RRRC4770
RRRC4780
RRRC4790
RRRC4800
RRRC4810
RRRC4820
RRRC4830
RRRC4840
RRRC4850
RRRC4860
RRRC4870
RRRC4880
RRRC4890
RRRC4900
RRRC4910
RRRC4920
RRRC4930
RRRC4940
RRRC4950
RRRC4960
RRRC4970
RRRC4980
RRRC4990
RRRC5000
RRRC5010
RRRC5020
RRRC5030
RRRC5040
RRRC5050
RRRC5060
RRRC5070
RRRC5080
RRRC5090
RRRC5100
RRRC5110

```

```

IF (M.EC.1) GO TO 332
IF (DISTB-GE-SMDIST) GO TO 335
332 ICGRP = M
335 CC CONTINUE
C IS CBSEVERATION ALREADY IN THE CLOSEST CLUSTER? IF YES, SKIP THIS
SECTION
IF (ICGRP.EQ.IDATAT(I)) GO TO 360
ICLD IS OLD CLUSTER ASSIGNMENT
346 ICCLC = ICATAT(I)
C INEW IS NEW CLUSTER ASSIGNMENT
INEW = IDGP
JFLAG = 1
RECALCULATE CLUSTER CENTERS
DC 348 J = 1, NVARS
SVCENT(ICLD,J) = (NISVT(ICLD)*SVCENT(ICLD,J)-DATA(I,J))/
1 (NISVT(ICLD)-1)
346 SVCENT(INEW,J) = (NISVT(INEW)*SVCENT(INEW,J)+DATA(I,J))/
1 (NISVT(INEW)+1)
C ADJUST NISVT
NISVT(ICLD) = NISVT(ICLD)-1
NISVT(INEW) = NISVT(INEW)+1
C ADJUST ICATAT
ICATAT(I) = IDGP
RECALCULATE WITHIN-CLUSTERS MATRIX FOR USE IN COMPUTING SCALED
EUCALC AND MAHALANOBIS DISTANCE
IF (IDIST.EC.0) GO TO 360
CALL WCALC (SVCENT,NISVT,NGPST,ICIR)
IF (IFINE.EC.1) RETURN
36C CC CONTINUE
C DCAE WITH MAJOR DO LOOP
4CC CC CONTINUE
C
C CALCULATE THE CRITERION
RECALCULATE SVCENT: ACCURACY MEASURE
LC 401 M=1,NGPST
DC 401 J=1,NVARS
SVCENT(M,J) = 0.
4C1 CC 402 I=1,NGBS
M = IDATAT(I)
DC 402 J=1,NVARS
SVCENT(M,J) = SVCENT(M,J) + DATA(I,J)/NISVT(M)
4C2 RECALCULATE WITHIN-CLUSTERS MATRIX AND CRITERIA VALUE BASED ON
NEW CLUSTER CENTER VALUES
ICIR = 2
IF (ICRIT.EC.1) ICIR = 1
CALL WCALC (SVCENT,NISVT,NGPST,ICIR)

```

```

IF (IFINE.EC.1) RETURN
CALL CRITCN (CRIT)
IF (IFINE.EC.1) RETURN
IF CRITERION BETTER THAN BEFORE? IF YES, THEN ANOTHER ITERATION:
IF NO, FINISH KMEANS
IF (CRIT.GE.BCRIT) GO TO 535

    ANOTHER ITERATION

PLT TEMPORARY VALUES INTO PERMANENT LOCATIONS
515 NGPS=NGPST
    CC 520 M = 1,NGPS
    NISV(M) = NISV(M)
    DC 520 J = 1,NVARS
    SVGEN (M,J) = SVCENT (M,J)
52C DC 530 I = 1,NOBS
53C ICATA(I) = IDATA(I)
    GC TO 200

    FINISH

JFLAG = C MEANS NO CHANGES HAVE BEEN MADE DURING THE LAST
ITERATION: ITERATION HAS CONVERGED
535 IF (JFLAG.EC.0) RETURN
JFLAG = 1 MEANS CHANGES HAVE BEEN MADE BUT THE CRITERION VALUE
GCT INCRSE: ITERATION HAS NOT CONVERGING
54C WRITE (6,94C) CRIT
    WRITE (6,942) BCRIT
    RECALCULATE WITHIN-CLUSTERS MATRIX AND RESET CRIT
    ICIR=2
    IF (ICIR.EC.1) ICIR=1
    CALL WCALC (SVGEN,NISV,NGPS,IDIR)
    IF (IFINE.EC.1) RETURN
    CRIT=BCRIT
    RETURN ('O ITERATION HAS NOT CONVERGING')
54C FCRMAT ('O THE CRITERION VALUE IS ',E12.6)
943 FCRMAT ('O THE BEST CRITERION VALUE IS ',E12.6)
943 END
SUBROUTINE ISWITCH (CRIT)

THIS SUBROUTINE CONSIDERS SWITCHING EACH OBSERVATION TO A
DIFFERENT CLUSTER. THE SWITCH IS MADE IFF A BETTER CRITERION
VALUE RESULTS.
THIS SUBROUTINE ALSO DEPENDS ON THE PARAMETER KTIME: IT DETERMINES
WHICH OBSERVATIONS ARE TO BE CONSIDERED. FOR COMPLETE EXPLANATION
OF THE KTIME PARAMETER, SEE THE PROGRAM DESCRIPTION.

```



```

C      KTIME = 9 MEANS CONSIDER ALL OBSERVATIONS
C      IF (KTIME.EC.9) GC TO 618
C      DO NOT CONSIDER SWITCHING THE OBSERVATION IF THE DISTANCE TO ITS
C      CLURRENT CLUSTER CENTER IS LT BIGCIS / KTIME
615  XVEC(J) = DATA(I,J)/YVEC(DISTA)
C      CALL DISTCE (XVEC,YVEC,DISTA)
C      IF (DISTA.LT.PSME) GO TO 691
61E  CCNTINUE
C      DO 690 MNEW=1,NGFS
C      IF (NISV(MOLD).EQ.1) GO TO 690
C      IF (MNEW.EC.MOLD) GC TO 690
C      IF (KFLAG.EC.1) GC TO 690
C      COMPUTE NEW SVCCEN BASED ON OBSERVATION BEING SWITCHED
C      DO 620 J=1,AVARS
C      SVCCEN(MNEW,J) = (NISV(MNEW)*SVCCEN(MNEW,J)+DATA(I,J))/
C      1 (NISV(MNEW)+1)
C      1 SVCCEN(MOLD,J) = (NISV(MOLD)*SVCCEN(MOLD,J)-DATA(I,J))/
C      1 (NISV(MOLD)-1)
62C  CCNTINUE
C      ADJUST NISV
C      NISVT(MNEW) = NISV(MNEW)+1
C      NISVT(MOLD) = NISV(MOLD)-1
C      COMPUTE WITHIN-CLUSTERS MATRIX AND NEW CRITERION VALUE
C      CALL WCALC (SVCCEN,NISVT,NGPST,ICIR)
C      IF (IFINE.EC.1) RETURN
C      CALL CRITON (TCRIT)
C      IF (IFINE.EC.1) RETURN
C      MAKE THE SWITCH IFF NEW CRIT IS BETTER
C      IF (TCRIT.GE.CRIT) GO TO 680
C      MAKE THE SWITCH
C      RESET FLAGS
C      JFLAG(M) = 1
C      JFLAG(MNEW) = 1
C      IFLAG=1
C      KFLAG = 1
C      ADJUST CRIT, IDATA, NISV, SVCCEN
C      TCRIT=TCRIT
C      ILATA(I) = MNEW
C      NISV(MNEW) = NISVT(MNEW)
C      NISV(MOLD) = NISVT(MOLD)
C      DO 630 J=1,AVARS
C      SVCCEN(MNEW,J) = SVCCEN(MNEW,J)
C      SVCCEN(MOLD,J) = SVCCEN(MOLD,J)
C      GC TO 690
C      SWITCH WAS ACT MADE; RESET NISVT, SVCCEN, TC VALUES PRESENT

```

```

RRRC6080
RRRC6090
RRRC6100
RRRC6110
RRRC6120
RRRC6130
RRRC6140
RRRC6150
RRRC6160
RRRC6170
RRRC6180
RRRC6190
RRRC6200
RRRC6210
RRRC6220
RRRC6230
RRRC6240
RRRC6250
RRRC6260
RRRC6270
RRRC6280
RRRC6290
RRRC6300
RRRC6310
RRRC6320
RRRC6330
RRRC6340
RRRC6350
RRRC6360
RRRC6370
RRRC6380
RRRC6390
RRRC6400
RRRC6410
RRRC6420
RRRC6430
RRRC6440
RRRC6450
RRRC6460
RRRC6470
RRRC6480
RRRC6490
RRRC6500
RRRC6510
RRRC6520
RRRC6530
RRRC6540
RRRC6550

```



```

C      BEFORE THE SWITCH WAS CONSIDERED
66C NISVT(MNEW) = NISV(MNEW)
67C NISVT(MOLD) = NISV(MOLD)
68C J=1, NVAR
69C SVCEN(MNEW,J) = SVCEN(MNEW,J)
70C SVCEN(MOLD,J) = SVCEN(MOLD,J)
71C CONTINUE
72C FINISH WITH CLUSTER M: IF NO SWITCHES HAVE BEEN MADE, SET
73C JFLAG(M) = 2 AND GO TO NEXT CLUSTER: IF SWITCHES HAVE BEEN MADE,
74C ADJUST LSTSV AND SET JFLAG(M) = C AND GO TO NEXT CLUSTER
75C IF (JFLAG(M).EQ.0) GO TO 699
76C JFLAG(M) = C
77C GO TO 700
78C JFLAG(M) = 2
79C CONTINUE
80C DCNE WITH ALL CLUSTERS: IFLAG = 1 MEANS SOME SWITCHES HAVE BEEN
81C MADE: GO BACK AND ITERATE
82C IF (IFLAG.EC.1) GO TO 600
83C
84C ALL DONE: ACCURATELY CALCULATE MEANS AND CRITERION AND CUTFUT
85C THE RESULTS
86C
87C WRITE (6,90C)
88C FCFORMAT (11, 'THE FINAL CLUSTER SOLUTION IS ')
89C RECALCULATE CLUSTER CENTERS, WITHIN-CLUSTERS MATRIX, AND CRITERION
90C VALUE
91C
92C DC 715 M=1,NGPS
93C DC 715 J=1,NVAR
94C DC 720 I=1,NBS
95C M = IDATA(I)
96C DC 720 J=1,NVAR
97C SVCEN(M,J) = SVCEN(M,J) + DATA(I,J)/NISV(M)
98C CALL WCALC (SVCEN,NISV,NGPS,IDIR)
99C IF (IFINE.EC.1) RETURN
100C CALL CRITON (CRIT)
101C IF (IFINE.EC.1) RETURN
102C CALL THE OUTPUT ROUTINE
103C CALL OUTPUT (CRIT)
104C WRITE (6,901)
105C FCFORMAT (10, 'END OF CLUSTER PROBLEM')
106C RETURN
107C ENCL
108C SUBROUTINE CUTPUT (CRIT)
109C
110C THIS SUBROUTINE PRINTS OUT THE CLUSTER SOLUTION
111C FOR EACH CLUSTER - THE CLUSTER SIZE

```

```

RRR06560
RRR06570
RRR06580
RRR06590
RRR06600
RRR06610
RRR06620
RRR06630
RRR06640
RRR06650
RRR06660
RRR06670
RRR06680
RRR06690
RRR06700
RRR06710
RRR06720
RRR06730
RRR06740
RRR06750
RRR06760
RRR06770
RRR06780
RRR06790
RRR06810
RRR06820
RRR06830
RRR06840
RRR06850
RRR06860
RRR06870
RRR06880
RRR06890
RRR06900
RRR06910
RRR06920
RRR06930
RRR06940
RRR06950
RRR06960
RRR06970
RRR06980
RRR06990
RRR07000
RRR07010
RRR07020
RRR07030

```



```

101 L=C
102 IF (L-EC.NISV(M)) GO TO 110
103 IF (ICATAT(L).LE.ICATAT(LL)) GO TO 101
104 LTEMP = IDATAT(L)
105 ICATAT(L) = IDATAT(LL)
106 ICATAT(LL) = LTEMP
107 IF (L-EC.1) GO TO 102
108 L=L-1
109 GC TO 102
110 CCNTINUE
111 WRITE OUT THE IDENTIFICATIONS
112 WRITE (6,904) (IDATAT(K),K=1,LSTNC)
113 WRITE (7,920) (IDATAT(K),K=1,LSTNC)
114 FCNMAT(1615)
115 CCNTINUE
116 WRITE OUT THE WITHIN-CLUSTERS MATRIX
117 WRITE (6,905)
118 DO 120 J=1,NVARS
119 WRITE (6,906) (W(K,J),K=1,J)
120 CCNTINUE
121 ICRT IS THE CRITERION CHOSEN BY THE USER
122 GC TO (15C,160,17C,170),ICRT
123 WRITE OUT TRACE W
124 WRITE (6,907) CRIT
125 RETURN
126 WRITE OUT DET W
127 ANCL LOG (DET T / DET W)
128 CALL UPRECT (NVARS,T,M)
129 DET = 1
130 DO 165 J=1,NVARS
131 DET = DET*(J,J)
132 WRITE (6,912) CRIT
133 CRIT = ALOG 10 ((DET**2)/(CRIT**2))
134 WRITE (6,905) CRIT
135 RETURN
136 FCF LARGEST FOOT AND PCTELLING'S TRACE CRITERIA, FIND EIGENVALUES
137 FOR 18-GW1=0,NVARS
138 DO 175 J=1,NVARS
139 DO 175 K=J,NVARS
140 BT(J,K) = B(J,K)
141 BT(K,J) = B(J,K)
142 CFT = 1.0 / CRIT
143 CALL UTISU (NVARS,BT,WFCFT)
144 CALL EIGN (NVARS,ET,EIG,VEC,IND)
145 IF (IND.NE.C) GO TO 180
146 WRITE OUT EIGENVALUES

```

```

RRR07500
RRR07510
RRR07520
RRR07530
RRR07540
RRR07550
RRR07560
RRR07570
RRR07580
RRR07590
RRR07600
RRR07610
RRR07620
RRR07630

```

```

RRR07640
RRR07650
RRR07660
RRR07670
RRR07680
RRR07690
RRR07700
RRR07710
RRR07720
RRR07730
RRR07740
RRR07750
RRR07760
RRR07770
RRR07780
RRR07790
RRR07800
RRR07810
RRR07820
RRR07830
RRR07840
RRR07850
RRR07860
RRR07870
RRR07880
RRR07890
RRR07900
RRR07910
RRR07920
RRR07930
RRR07940
RRR07950

```

```

C
WRITE (6,91C) (EIG(J),J=1,NVARS)
IF (ICRIT.EC.4) GC TO 178
WRITE OUT LARGEST FCCT
RETURN
C
WRITE OUT SUM OF EIGENVALUES
178 WRITE (6,914) CRIT
RETURN
C
ESCAPE ROUTE
18C IFINE = 1
WRITE (6,911) IND
RETURN
C
FCMAT STATEMENTS
9CC
FCMAT (,'OCCLUSTER',12)
FCMAT (,'SIZE',13)
9C1 FCMAT (,'CENTER',15(F9.2,1X),/,18X,5(F5.2,1X))
9C2 FCMAT (,'THE OBSERVATIONS ARE :')
9C3 FCMAT (,'X',57,/,5X,517)
9C4 FCMAT (,'1',THE WITHIN GROUP MATRIX IS *)
9C5 FCMAT (,'1',IC(F10.3,1X))
9C6 FCMAT (,'0',THE TRACE OF THE WITHIN-CLUSTERS MATRIX IS ',E12.6)
9C7 FCMAT (,'0',THE DETERMINANT T IS ',E12.6)
9C8 FCMAT (,'0',LOG (DET T / DET W) = ',E12.6)
9C9 FCMAT (,'0',THE EIGENVALUES OF W(INVERSE)*E ARE ',10F8.3)
9C10 FCMAT (,'0',THE EIGENVALUES OF THE WITHIN-CLUSTERS MATRIX IS ',E12.6)
9C11 FCMAT (,'1',ILL-CONDITIONED W MATRIX -- IND IS ',13)
9C12 FCMAT (,'0',THE DETERMINANT OF THE WITHIN-CLUSTERS MATRIX IS ',E12.6)
9C13 FCMAT (,'0',THE LARGEST ROOT IS ',E12.6)
9C14 FCMAT (,'0',THE TRACE OF W(INV)*B IS ',E12.6)
ENC
SUBROUTINE DISTCE (XVECS,YVECS,DIST)
THIS SUBROUTINE CALCULATES EUCLIDIAN, WEIGHTED EUCLIDIAN, CR
MAHALANOBIS DISTANCE (DEPENDING ON IDIST).
CC
CC
CC
COMMON NOBS,NVARS,NGPS,ICRIT,NCSTAN,IDIST,IFINE,KTIME,ICENT(6C0),
IGATA(6C0,20),T(2C,20),B(2C,20),H(2C,20),WFC(2C,20),SVCE(20,2C),
21 ICATA(6C0),ALSV(20),VMEAN(20),SC(2C),XVEC(2C),VVEC(2C),E(2C,2C),
3 NISSVT(20),SVCENT(20,20),IDATAT(6C0),VEC(20,20),EIG(20),
DIMENSION X(20),XVECS(1),YVECS(1)
ICIR = IDIST+1
GC TO (10C,20C,30C),IDIR
WRITE (6,90C)
90C IFINE = 1
RETURN
C
RRRC756C
RRRC757C
RRRC758C
RRRC759C
RRRC800C
RRRC801C
RRRC802C
RRRC803C
RRRC804C
RRRC805C
RRRC806C
RRRC807C
RRRC808C
RRRC809C
RRRC810C
RRRC811C
RRRC812C
RRRC814C
RRRC817C
RRRC818C
RRRC819C
RRRC820C
RRRC821C
RRRC822C
RRRC823C
RRRC824C
RRRC825C
RRRC826C
RRRC827C
RRRC828C
RRRC829C
RRRC830C
RRRC831C
RRRC832C
RRRC833C
RRRC834C
RRRC835C
RRRC836C
RRRC837C
RRRC838C
RRRC839C
RRRC840C
RRRC841C
RRRC842C
RRRC843C

```



```

C 200 C E T = 1.
C 205 C C J=1, N V A R S
C 205 C E T = D E T * W F C T ( J , J )
C 205 C R I T = C E T
C 205 C R E T U R N
C
C 300 C C A L C U L A T E L A R G E S T R O O T A S L ( I N V ) * R * L * ( I N V ) W H E R E L I S T H E
C 300 C C H E S K Y F A C T O R O F W
C 300 C C R I T E R I O N I S R E C I P R O C A L O F L A R G E S T R O O T
C 300 C C
C 305 C C J=1, N V A R S
C 305 C C K=J, N V A R S
C 305 C C B T ( J , K ) = B ( J , K )
C 305 C C B T ( K , J ) = B T ( J , K )
C 305 C C C A L L U T I S U I ( N V A R S , B T , W F C T )
C 305 C C C A L L E I G N ( N V A R S , E T , E I G , V E C , I N D )
C 305 C C C R I T = 1.0 / E I G ( 1 )
C 305 C C R E T U R N
C
C 400 C C A L C U L A T E H C T E L L I N G ' S T R A C E A S S U M O F D I A G O N A L E L E M E N T S O F
C 400 C C L ( I N V ) * R * L * ( I N V )
C 400 C C C R I T E R I O N I S R E C I P R O C A L O F T H I S S U M
C 400 C C
C 405 C C J=1, N V A R S
C 405 C C K=J, N V A R S
C 405 C C B T ( J , K ) = B ( J , K )
C 405 C C B T ( K , J ) = B T ( J , K )
C 405 C C C A L L U T I S U I ( N V A R S , B T , W F C T )
C 405 C C C R I T = 0.
C 405 C C C R I T = C R I T + B T ( J , J )
C 405 C C C R I T = 1.0 / C R I T
C 405 C C R E T U R N
C 405 C C E N C
C 405 C C S L E R O U T I N E W C A L C ( S V , N I , N G P , I D I R )
C
C 500 C T H I S S U B R O U T I N E C A L C U L A T E S T H E W I T H I N - C E L L S M A T R I X A N D , I F
C 500 C N E C E S S A R Y , T H E C H E S K Y F A C T O R O F T H E W I T H I N - C E L L S M A T R I X
C 500 C
C 505 C C C M M C N N O P S , N V A R S , N G P S , I C R I T , N O S T A N , I D I S T , I F I N E , K T I M E , I C E N T ( 6 C C ) ,
C 505 C C 1 D A T A ( 6 C C , 20 ) , T ( 20 , 20 ) , B ( 20 , 20 ) , W ( 20 , 20 ) , W F C T ( 20 , 20 ) , S V C E N ( 20 , 20 ) ,
C 505 C C 2 I C A T A ( 6 C C ) , N I S V ( 20 ) , V M E A N ( 20 ) , S C ( 20 ) , X V E C ( 20 ) , V E C ( 20 ) , E T ( 20 , 20 ) ,
C 505 C C 3 N I S V T ( 20 ) , S V C E N T ( 20 , 20 ) , I D A T A T ( 6 C C ) , V E C ( 20 , 20 ) , E I G ( 20 ) ,
C 505 C C C I M E N S I O N S V ( 20 , 20 ) , N I ( 20 )
C
C 600 C C A L C U L A T E B A N D T H E N W = T - B
C 600 C B = S U M N I S \ ( M ) * S V C E N ( M ) * 2

```

```

RRRC8520C
RRRC8530C
RRRC8540C
RRRC8550C
RRRC8560C
RRRC8570C
RRRC8580C
RRRC8590C
RRRC8600C
RRRC8610C
RRRC8620C
RRRC8630C
RRRC8640C
RRRC8650C
RRRC8660C
RRRC8670C
RRRC8680C
RRRC8690C
RRRC8700C
RRRC8710C
RRRC8720C
RRRC8730C
RRRC8740C
RRRC8750C
RRRC8760C
RRRC8770C
RRRC8780C
RRRC8790C
RRRC8800C
RRRC8810C
RRRC8820C
RRRC8830C
RRRC8840C
RRRC8850C
RRRC8860C
RRRC8870C
RRRC8880C
RRRC8890C
RRRC8900C
RRRC8910C
RRRC8920C
RRRC8930C
RRRC8940C
RRRC8950C
RRRC8960C
RRRC8970C
RRRC8980C
RRRC8990C
RRRC9000C
RRRC9010C
RRRC9020C
RRRC9030C
RRRC9040C
RRRC9050C
RRRC9060C
RRRC9070C
RRRC9080C
RRRC9090C
RRRC9100C
RRRC9110C
RRRC9120C
RRRC9130C
RRRC9140C
RRRC9150C
RRRC9160C
RRRC9170C
RRRC9180C
RRRC9190C
RRRC9200C
RRRC9210C
RRRC9220C
RRRC9230C
RRRC9240C
RRRC9250C
RRRC9260C
RRRC9270C
RRRC9280C
RRRC9290C
RRRC9300C
RRRC9310C
RRRC9320C
RRRC9330C
RRRC9340C
RRRC9350C
RRRC9360C
RRRC9370C
RRRC9380C
RRRC9390C

```

```

C
100 J=1,NVARS
100 K=J,NVARS
100 E(J,K) = 0.
105 N=1,NGP
105 J=1,NVARS
105 K=J,NVARS
105 B(J,K) = SV(M,J) * SV(M,K) + ET(J,K) * NI(M)
105 E(J,K) = B(J,K) + ET(J,K) * NI(M)
110 J=1,NVARS
110 K=J,NVARS
110 W(J,K) = T(J,K) - E(J,K)
C
C CALCULATE CHOLESKY FACTOR OF W IF ICIR = 2
C
115 GO TO (120,115),ICIR
115 CC 116 J=1,NVARS
116 CC 116 K=J,NVARS
116 WFACT(J,K) = W(J,K)
120 CALL UPFECT (NVARS,WFACT,M)
120 IF (M.NE.0) GO TO 125
120 RETURN
125 IFINE=1
125 WRITE (6,90C)
90C FORMAT ('1 THE WITHIN GROUPS MATRIX IS SINGULAR ')
90C RETURN
C
C FUNCTION URANC(IRANC)
C
THIS FUNCTION CALCULATES UNIFORMLY DISTRIBUTED RANCC# NUMBERS.
BETWEEN 0 AND 1
3*19 CONGRUENTIAL UNIFORM RANCC# NUMBER GENERATOR
C
IRAND = IRAND*1162261467
IF (IRAND.GT.0) GO TO 3
IRAND = -IRAND
URAND = FLOAT(IRAND)*0.4656612873E-9
RETURN
END
C
C SUBROUTINE LPRFCT(N,A,M)
C
REPLACE UPPER TRIANGLE OF A SQUARE POSITIVE DEFINITE MATRIX A
BY ITS CHOLESKI FACTOR
C
C DIMENSION A(20,20)
C CLEAR THE ERROR INDICATOR
M=0
NI=N-1

```

```

RRRC540C
RRRC541C
RRRC542C
RRRC543C
RRRC544C
RRRC545C
RRRC546C
RRRC547C
RRRC548C
RRRC549C
RRRC550C
RRRC551C
RRRC552C
RRRC553C
RRRC554C
RRRC555C
RRRC556C
RRRC557C
RRRC558C
RRRC559C
RRRC560C
RRRC561C
RRRC562C
RRRC563C
RRRC564C
RRRC565C
RRRC566C
RRRC567C
RRRC568C
RRRC569C
RRRC570C
RRRC571C
RRRC572C
RRRC573C
RRRC574C
RRRC575C
RRRC576C
RRRC577C
RRRC578C
RRRC579C
RRRC580C
RRRC581C
RRRC582C
RRRC583C
RRRC584C
RRRC585C
RRRC586C
RRRC587C

```

```

100 IF(N1) 23C,10C,10C
101 DC 220 K=1,N
102 ANK=A(K,K)
103 IF(K .EC. 1) GO TC 12C
104 DC 110 J=2,K
105 ANK=AKK-A(J-1,K)*2
106 IF(A(K,K)) 14C,14C,13C
107 IN THE CASE OF A COVARIANCE MATRIX ANK/A(K,K) IS 1-F**2 WHERE
108 R IS MULTICORRELATION OF VARIABLE K WITH ALL PRECEDING VARIABLES
109 IF(A(K,K)/A(K,K) .GE. .001) GO TC 150
110 DC 14C M=K
111 ANK=0
112 ANK=SQRT(A(K,K))
113 A(K,K)=ANK
114 IF(K .EC. N) GO TC 23C
115 C
116 DC 220 I=K,N1
117 ANI=A(K,I+1)
118 IF(K .EC. 1) GO TC 190
119 DC 180 J=2,K
120 ANI=AKI-A(J-1,K)*A(J-1,I+1)
121 IF(A(K,I+1))=0.0
122 GO TO 220
123 A(K,I+1)=AKI/ANK
124 CCNT=CCNT+1
125 RETURN
126 ENCL
127 SUBROUTINE LTIRT(M,N,S,B)
128 C
129 INVERSE OF UPPER (S) TRANSPOSED TIMES RECTANGLE B TRANSPOSED.
130 C
131 DIMENSION S(20,20),B(1,20)
132 DC 130 J=1,N
133 DC 130 I=1,M
134 SUM=0.0
135 IF (S(I,I)) 90,130,90
136 SLA = B(J,I)
137 IF(I=1) 120,120,100
138 IF(IM1) K=1,IM1
139 CC 110 K=1,IM1
140 SLA = SUM-S(K,I)*B(J,K)
141 SLA = SUM/S(I,I)
142 B(J,I) = SLA
143 RETURN
144 ENCL
145 SUBROUTINE LTISUI (N,A,B)
146 C

```



```

RRR1C840
RRR1C850
RRR1C86C
RRR1C87C
RRR1C880
RRR1C89C
RRR1C90C
RRR1C910
RRR1C920
RRR1C930
RRR1C950
RRR1C96C
RRR1C97C
RRR1C980
RRR1C990
RRR1C100
RRR1C110
RRR1C120
RRR1C130
RRR1C140
RRR1C150
RRR1C160
RRR1C170
RRR1C180
RRR1C190
RRR1C200
RRR1C210
RRR1C220
RRR1C230
RRR1C240
RRR1C250
RRR1C260
RRR1C270
RRR1C280
RRR1C290
RRR1C300
RRR1C310

```

```

C      ESCLIVALENCE (IPOS(1), GAMMA(1)), (IVPOS(1), PETA(1)),
1      I(CRD(1), PETA(1))
C      N=NN
C      RESET ERROR RETURN INDICATOR
C      INC=0
C      IF(N .EC. 0) GO TO 560
C      N=N-1
C      N=N-2
C      COMPUTE THE TRACE AND EUCLIDIAN NORM OF THE INPUT MATRIX
C      LATER CHECK AGAINST SUM AND SUM OF SQUARES OF EIGENVALUES
C      ENCRM=0.
C      TRACE=0.
C      DO 110 J=1, N
C      DO 100 I=J, N
C      ENCRM=ENCRM+A(I, J)**2
C      TRACE=TRACE+A(I, J)
C      ENCRM=ENCRM-.5*A(J, J)**2
C      GAMMA(1)=A(1, 1)
C      IF (N2) 28C, 270, 12C
C      DO 26C NR=1, N2
C      E=A(NR+1, NR)
C      S=C
C      DO 130 I=NR, N2
C      S=S+A(I+2, NR)**2
C      PREPARE FOR POSSIBLE BYPASS OF TRANSFORMATION
C      A(NR+1, NR)=C
C      IF (S) 25C, 250, 140
C      S=S+B*B
C      SGN=+1
C      IF (B) 150, 16C, 160
C      SGN=-1
C      SCRTS=SCRT(S)
C      S=SGN/(SCRTS+SCRTS)
C      TEMP=SCRT(.5+B*C)
C      W(NR)=TEMP
C      A(NR+1, NR)=TEMP
C      L=L/TEMP
C      B=-SGN*SCRTS
C      IS FACTOR OF PROPORTIONALITY. NOW COMPUTE AND SAVE A VECTOR.
C      EXTRA SINGLY SUBSCRIPTED W VECTOR USED FOR SPEED.
C      DO 170 I=NR, N2
C      TEMP=D*A(I+2, NR)
C      W(I+1)=TEMP
C      A(I+2, NR)=TEMP
C      PREMULTIPLY VECTOR W BY MATRIX A TO OBTAIN F VECTOR.
C      SIMULTANEOUSLY ACCUMULATE DCT PRODUCT WP, (THE SCALAR K)
C      W*WAM=0.

```

```

RRRI113200
RRRI113300
RRRI113400
RRRI113500
RRRI113600
RRRI113700
RRRI113800
RRRI113900
RRRI114000
RRRI114100
RRRI114200
RRRI114300
RRRI114400
RRRI114500
RRRI114600
RRRI114700
RRRI114800
RRRI114900
RRRI115000
RRRI115100
RRRI115200
RRRI115300
RRRI115400
RRRI115500
RRRI115600
RRRI115700
RRRI115800
RRRI115900
RRRI116000
RRRI116100
RRRI116200
RRRI116300
RRRI116400
RRRI116500
RRRI116600
RRRI116700
RRRI116800
RRRI116900
RRRI117000
RRRI117100
RRRI117200
RRRI117300
RRRI117400
RRRI117500
RRRI117600
RRRI117700
RRRI117800
RRRI117900

```

```

CC 220 I=NR,N1
SLM=0.
CC 180 J=NR,I
SLM=SLM+A(I+1,J+1)*W(J)
21C I=I+1
IF(N1-I) 210,190,190
19C DC 200 J=I,N1
22C SLM=SUM+A(J+1,I+1)*W(J)
23C P(I)=SUM
P(I)=W(AH+SLM*W(I))
F VECTOR ANI SCALAR K NOW STORED. NEXT COMPLETE Q VECTOR
CC 230 I=NR,N1
Q(I)=P(I)-W(AH*W(I))
C 23C NCM FORM PAF MATRIX, REQUIRED PART
DC 240 J=NR,N1
QJ=Q(J)
H=H(J)
DC 240 I=J,N1
A(I+1,J+1)=A(I+1,J+1)-2.*(W(I)*QJ+HJ*Q(I))
24C BETA(NR)=B
25C BETASQ(NR)=E*B
26C GAMMA(NR+1)=A(NR+1,NR+1)
27C E=A(N,N-1)
BETA(N-1)=B
BETASQ(N-1)=B*B
GAMMA(N)=A(A,N)
BETASQ(N)=0.
C 28C ACJGIN AN ICENTITY MATRIX TC BE PCSTMULTIPLIED BY ACTIATICS.
DC 290 I=1,N
CC 300 J=1,N
VEC(I,J)=0.
30C VEC(I,I)=1.
SLM=0.
AFAS=1
AC TO 400
31C SLM=SUM+SHIFT
CCSA=1.
G=GAMMA(1)-SHIFT
PF=G
FFES=PP*PP+ETASQ(1)
PFER=SQRT(PFES)
CC 370 J=1,N
CCSAP=CCSA
IF(PFBS .NE. 0.) GC TC 320
SINA2=0.
CCSA=1.

```

```

RRRI11800C
RRRI11810C
RRRI11820C
RRRI11830C
RRRI11840C
RRRI11850C
RRRI11860C
RRRI11870C
RRRI11880C
RRRI11890C
RRRI11900C
RRRI11910C
RRRI11920C
RRRI11930C
RRRI11940C
RRRI11950C
RRRI11960C
RRRI11970C
RRRI11980C
RRRI11990C
RRRI12000C
RRRI12010C
RRRI12020C
RRRI12030C
RRRI12040C
RRRI12050C
RRRI12060C
RRRI12070C
RRRI12080C
RRRI12090C
RRRI12100C
RRRI12110C
RRRI12120C
RRRI12130C
RRRI12140C
RRRI12150C
RRRI12160C
RRRI12170C
RRRI12180C
RRRI12190C
RRRI12200C
RRRI12210C
RRRI12220C
RRRI12230C
RRRI12240C
RRRI12250C
RRRI12260C
RRRI12270C

```

```

GC TO 350
SINA=BETA(J)/PPBR
SINA2=BETASQ(J)/PPBS
CCSA=FP/PPBR
CCLSMULTIPLY IDENTITY BY P-TRANSPCSE MATRIX
NT=J+NPAS
IF(NT.EQ. N) NT=N
GC 340 I=1,NT
TEMP=CCSA*VEC(I,J)+SINA*VEC(I,J+1)
VEC(I,J+1)=-SINA*VEC(I,J)+CCSA*VEC(I,J+1)
C 34C VEC(I,J)=TEMP
C 35C CIA=GAMMA(J+1)-SHIFT
L=SINA2*(G+CIA)
GAMMA(J)=G+L
G=CIA-U
PF=DIA*CCSA-SINA*CCSAP*BETA(J)
IF(J.NE. N) GO TC 360
BETA(J)=SINA*PP
BETASQ(J)=SINA2*PP*PP
GC TO 380
PFES=PP*PP+FEETASQ(J+1)
PFER=SQRT(PFES)
FEETASQ(J)=SINA*PPBR
BETASQ(J)=SINA2*PPBS
C 36C PFES=PP*PP+FEETASQ(J+1)
C 37C FEETASQ(J)=SINA*PPBR
C 38C BETASQ(J)=SINA2*PPBS
C TEST FOR CONVERGENCE OF LAST DIAGONAL ELEMENT
NFAS=NPAS+1
IF(BETASQ(M).GT. 1.E-21) GC TO 41C
EIG(M+1)=GAMMA(M+1)+SUM
BETASQ(M)=0.
BETASQ(M)=C.
M=M-1
IF(M.EQ. 0) GO TC 43C
IF(BETASQ(M).LE. 1.E-21) GO TO 350
TAKE ROOT OF CORNER 2 BY 2 NEAREST TO LOWER DIAGONAL IN VALUE
C 41C AS ESTIMATE OF EIGENVALUE TO USE FOR SHIFT
C 42C A2=GAMMA(M+1)
C 43C R1=.5*A2
C 44C R12=R1+R2
C 45C TEMP=SQRT(DIF*DIF+BETASQ(M))
C 46C R12=R12+TEMP
C 47C R2=ABS(A2-R1)-ABS(A2-R2)
C 48C IF(DIF.LT. 0.) GC TO 420
C 49C SHIFT=R2
GC TO 310

```

```

RRR122290C
RRR122300C
RRR122310C
RRR122320C
RRR122330C
RRR122340C
RRR122350C
RRR122360C
RRR122370C
RRR122380C
RRR122390C
RRR122400C
RRR122410C
RRR122420C
RRR122430C
RRR122440C
RRR122450C
RRR122460C
RRR122470C
RRR122480C
RRR122490C
RRR122500C
RRR122510C
RRR122520C
RRR122530C
RRR122540C
RRR122550C
RRR122560C
RRR122570C
RRR122580C
RRR122590C
RRR122600C
RRR122610C
RRR122620C
RRR122630C
RRR122640C
RRR122650C
RRR122660C
RRR122670C
RRR122680C
RRR122690C
RRR122700C
RRR122710C
RRR122720C
RRR122730C
RRR122740C
RRR122750C

```

```

42C SHIFT=R10
43C GO TO 310
C INITIALIZE AUXILIARY TABLES REQUIRED FOR REARRANGING THE VECTORS
  DC 440 J=1,N
  IFCSV(J)=J
  IVPQS(J)=J
44C USE A TRANSPOSITION SORT TO ORDER THE EIGENVALUES
  M=N
  GC TO 470
  CC 460 J=1,P
  IF(EIG(J)) .CE. EIG(J+1)) GO TO 46C
  TEMP=EIG(J)
  EIG(J)=EIG(J+1)
  EIG(J+1)=TEMP
  ICRD(J)=ICRD(J+1)
  ICRD(J+1)=ICRD(J)
  CC CONTINUE
46C M=M-1
47C IF(M .NE. 0) GO TO 45C
  IF(N1 .EQ. 1) GO TO 500
  CC 490 L=1,N1
  NV=IPCSV(L)
  NF=IPCSV(NV)
  IF(NP .EQ. L) GO TO 490
  LV=IVPOS(L)
  IVPQS(NP)=LV
  IPQS(LV)=NF
  CC 480 I=1,N
  TEMP=VEC(I,L)
  VEC(I,L)=VEC(NP,L)
  VEC(NP,L)=TEMP
  CC CONTINUE
  ESUM=0.
  BACK TRANSFORM THE VECTORS OF THE TRIPLE DIAGONAL MATRIX
  CC 550 NRR=1,N
  K=N1
  K=K-1
  IF(K .LE. 0) GO TO 540
  CC 520 I=K,N1
  SLP=SUM+VEC(I+1,NRR)+A(I+1,K)
  SLP=SLP+SUM
  CC 530 I=K,N1
  VEC(I+1,NRR)=VEC(I+1,NRR)-SUM+A(I+1,K)

```

RRR1276C
 RRR1277C
 RRR1278C
 RRR1279C
 RRR1280C
 RRR1281C
 RRR1282C
 RRR1283C
 RRR1284C

```

54C GC TO 51C +EIG(NRR)
55C EIG(NRR) +EIG(NRR) **2
    TEMP=ABS(TRACE-ESUM)+TEMP) -TEMP .NE. 0.) INC=INC+1
    IF((ABS(TRACE-ESUM)+TEMP) -TEMP .NE. 0.) INC=INC+2
    IF((ABS(TRACE-ESUM)+TEMP) -TEMP .NE. 0.) INC=INC+2
56C RETURN
    END
  
```

```

Appendix C:
RUN NAME
VARIABLE LIST
INPLT MEDIUM
INPLT MCRMAT
N OF CASES
VAR LABELS

PRINCIPAL CCMPONENT ANALYSIS OF TANKS
NATIONS,X1 TC X10
CARD
FIXED(1X,F2.0,5X,8F8.3/8),2F8.3)
24 NATIONS TYPE OF TANK/
X1 WEIGHT/X2 LENGTH/X3 WIDTH/X4 HEIGHT/X5 ROAD SPEED/
X6 TRENCF CRCSING/X7 GRCLNC PRESSURE/X8 MAX ARMAMENT/
X5 GROUND CLEARANCE/X10 FCWR TO ENGINE RATIC/
VARIABLES=X1 TO X10/TYPE=PAI/NFACTCRS=10/
8,11
FACTCR
OPTICS
STATISTICS
REAC INFLT DATA
DATA
.
.
.
.
FINISH

```

```

Appendix D:
RUN NAME LIST
VARIABLES
N CFCASES
INPLT FCRMAT
RECCE
VAR LABELS
DISCRIMINANT
CPTICAS
STATISTICS
REAL INFLT DATA
DATA
.
.
.
.
.
FINISH

TANK DISCRIMINANT
NATIONS,X1 TC X10
24
CAPD
FIXED(IX,F2-C,5X,8F8.3/6X,2F8.3)
NATIONS(11,13=1),(15,16,17,24,25,26,30,31,32,33=2)
(19,20,21,22,23,28,29,=3)(14,18,27,34=4)
NATIONS TYPE OF TANK/
X1 WEIGHT/X2 LENGTH/X3 WIDTH/X4 HEIGHT/X5 ROAD SPEED/
X6 TRENCH CROSSING/X7 GROUND PRESSURE/X8 MAX ARMAMENT/
X9 TRENCH CLEARANCE/X10 POWER TO ENGINE RATIC/
GRUPS=NATIONS(1,4)/VARIABLES=X1 TC X10/
ANALYSIS=X1 TO X10/METH-CC=MAHAL/
5,6,7,9,11,12
ALL
CPTICAS
STATISTICS
REAL INFLT DATA
DATA
.
.
.
.
.
FINISH

```


BIBLIOGRAPHY

1. Aiken, J. W., "Development of Cluster Analysis for Student Opinion Data", Master's Thesis, Naval Postgraduate School, Monterey, CA, 1979.
2. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1958.
3. Anderberg, M. R., Cluster Analysis for Applications, Academic Press, 1973.
4. Barr, D. R., and Richards, F. R., Utility Assessment Methodology (Report on Relative Utility Score of the AN-TPQ/27). System Exploration Inc., Monterey, CA, 1980.
5. Eisenheis, R. A., and Avery, R. B., Discriminant Analysis and Classification Procedures, Lexington Books, 1972.
6. Friedman, H. P., and Rubin, J., "On Some Invariant Criteria for Grouping Data", *Journal of the American Statistical Association*, Vol. 62: 320, Dec. 1967.
7. Giri, N. C., Multivariate Statistical Inference, Academic Press, 1977.
8. Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations, Wiley, 1977.
9. Green, P. E., and Carroll, J. D., Mathematical Tools for Applied Multivariate Analysis, Academic Press, 1976.
10. Hartigan, J. A., Clustering Algorithms, Wiley, 1975.
11. Kandell, M. G., The Advanced Theory of Statistics, Vol. III, Charles Griffin and Company, 1947.
12. Keeney, R. L., and Raiffa, H., Decisions with Multiple Objectives, Wiley, 1976.
13. Klecka, W. R., "Discriminant Analysis", SPSS (Statistical Package for the Social Sciences), McGraw Hill, 1975.
14. Kim, J. O., "Factor Analysis", SPSS (Statistical Package for the Social Sciences), McGraw Hill, 1975.

15. MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations", Paper presented at Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, 1965-1966.
16. McRae, D. J., Clustering Multivariate Observations, Doctoral Dissertation, University of North Carolina, Chapel Hill, N.C., 1973.
17. Morrison, D. F., Multivariate Analysis: Technique for Educational and Psychological Research, Wiley, 1971.
18. The IMSL Library, Vol. 3, IMSL, Inc., 1979.

INITIAL DISTRIBUTION LIST

	No. Copies
1. Defense Technical Information Center (DTIC) Cameron Station Alexandria, Virginia 22314	2
2. Library, Code 0142 Naval Postgraduate School Monterey, California 93940	2
3. Department Chairman, Code 55 Department of Operations Research Naval Postgraduate School Monterey, California 93940	4
4. Professor F. R. Richards, Code 55Rh (Thesis Advisor) Department of Operations Research Naval Postgraduate School Monterey, California 93940	2
5. Professor D. R. Barr, Code 55Bn (Second Reader) Department of Operations Research Naval Postgraduate School Monterey, California 93940	1
6. LTC Jin Ki Lee, ROK Army (Student) 6-601, Gun-In Apt, Dong-Bing-Go Dong Yong-San, Seoul Korea	5

LMED
-8